Modelling allergenic risk

Birot, Sophie

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

Modelling allergenic risk

Sophie BIROT

Kongens Lyngby 2016

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Richard Petersens Plads, building 324,
2800 Kongens Lyngby, Denmark
Phone +45 4525 3031
compute@compute.dtu.dk
www.compute.dtu.dk

Summary (English)

Up to 20 million Europeans suffer from food allergies. Due to the lack of knowledge about why food allergies developed or how to protect allergic consumers from the offending food, food allergy management is mainly based on food allergens avoidance. The iFAAM project (Integrated approaches to Food Allergen and Allergy Management) aims at developing strategies for food allergies based on evidences.

Especially, food allergen risk assessment helps food producers or authorities to make decisions on withdrawing a food product from the market or adding more information on the label when allergen presence is unintended. The risk assessment method has three different kinds of input. The exposure is calculated from the product consumption and the allergen contamination in the food product. The exposure is then compared to the thresholds to which allergic individuals react in order to calculate the chance of allergic reaction in the population.

In allergen risk assessment, the emphasis was on the threshold data, and no effort was made on consumption data. Moreover, no pan-European consumption data suitable for allergen risk assessment are available. A procedure for grouping food products automatically across countries is proposed. Thus, the allergen risk assessment can be performed cross-nationally and for the correct food group.

Then the two probabilistic risk assessment methods usually used were reviewed and compared. First order Monte-Carlo simulations are used in one method

[14], whereas the other one combines second order Monte-Carlo simulations with Bayesian inferences [13]. An alternative method using second order Monte-Carlo simulations was proposed to take into account the uncertainty from the inputs. The uncertainty propagation from the inputs to the risk of allergic reaction was also evaluated for all the methods using uncertainty analysis [11].

The recommended approach for the allergen risk assessment was implemented in a Shiny application with the R software. Thus, allergen risk assessment can be performed easily by non-statisticians with the interactive application.

Summary (Danish)

Op til 20 millioner europæere lider af fødevareallergi. På grund af manglende viden om udviklingen af fødevareallergi samt beskyttelse af allergiske forbrugere mod den fødevare de ikke kan tåle, så er kontrol af fødevareallergi hovedsagligt bygget på at undgå allergenerne i fødevare. Projektet iFAAM (Integrated approaches to Food Allergen and Allergy Management) har til hensigt at udvikle strategier for fødevareallergi baseret på beviser.

Fødevareallergeners risikovurdering hjælper fødevareproducenter eller myndigheder til at tage beslutninger om at trække et produkt tilbage fra markedet eller tilføje mere information til etiketten når et allergens tilstedeværelse er utilsigtet. Metoden for risikovurderingen har tre forskellige typer af input. Eksponeringen er udregnet ud fra produktindtag og mængden af forurening fra allergenet i fødevareproduktet. Eksponeringen er derefter sammenlignet med grænseværdier for hvornår allergiske individer reagerer for at udregne chancen for en allergisk reaktion i befolkningen.

Inden for allergenrisikovurderinger blev der lagt vægt på grænseværdidata, og der blev ikke arbejdet med forbrugsdata. Derudover, er intet paneuropæisk forbrugsdata tilgængeligt, der kan bruges til risikovurdering. En procedure til at gruppere fødevarer automatisk på tværs af lande er foreslået. Dermed kan allergenrisikovurdering udføres på tværs af landegrænser og for den korrekte fødevaregruppe.

Derefter blev de to probabilistiske risikovurderingsmetoder, som normalt bliver brugt, gennemgået og sammenlignet. Førsteordens Monte-Carlo simuleringer er brugt i en metode [14], hvor den anden kombinerer andenordens Monte-Carlo

simuleringer med Bayesiansk inferens [13]. En alternativ metode som bruger andenordens Monte-Carlo simuleringer blev foreslået, så den tager højde for usikkerhed fra inputs. Formering af usikkerhed fra inputs til risikoen for en allergisk reaktion blev også evalueret for alle metoderne ved at bruge usikkerhedsanalyse.

Den foreslåede fremgangsmåde for allergenrisikovurdering blev implementeret i en Shiny applikation med softwaren R. Dermed kan allergenrisikovurdering udføres nemt af ikke-statistikere med en interaktiv applikation.

Preface

This thesis was prepared at the Section of Statistics and Data Analysis of the department of Applied Mathematics and Compute Science (DTU Compute) at the Technical University of Denmark (DTU), in partial fulfilment of the requirements for acuiring the Ph.D. degree in Applied Mathematics. The project was funded by the iFAAM European project (Integrated Approaches to Food Allergen and Allergy Risk Management), Grant Agreement No. 322147.

The thesis deals with with modelling allergenic risk within the the iFAAM project. Food allergies are a growing concern in Europe and some mathematical and statistical methods are developed to assess the risk related to food allergies. The main focuses are on developing statistical procedure for grouping food consumption for allergen risk assessment and on reviewing and improving the models for food allergen risk assessment.

The thesis consists of three research paper included in the thesis, a Shiny application documented by it chapter. An introductory part provides an overview of the thesis and background information of the iFAAM project. Some investigations not detailed in the papers are also detailed in the thesis.

Lyngby, 31-October-2016

Sophie BIROT

List of contributions

Manuscripts

[3.6] **Birot, S.**, Madsen C.B., Kruizinga A.G., Christensen T., Crépet A., Brockhoff, P.B. *Grouping food consumption data for use in food allergen risk assessment.* (submitted to Journal of Food Composition and Analysis)

[3.7] **Birot, S.**, Madsen C.B., Kruizinga A.G., Crépet A., Christensen T., Brockhoff, P.B. *Combining food consumption data from different countries for creating food groups for allergen risk assessment (in Europe).* (submitted to Nutrients)

[4.2] **Birot, S.**, Crépet A., Remington B., Madsen C.B., Kruizinga A.G., Brockhoff, P.B. *Allergen probabilistic risk assessment modelling: existing model comparison and proposition of an alternative frequentist approach to account for uncertainty).* (draft paper, to be submitted to Risk Analysis)

Presentations

Half-year iFAAM meetings

One or two presentations were given at each half-year iFAAM project meeting to the work package partners. Some investigations carried out were presented and some methodological point were discussed. The list of project meetings can be found below:

- October 2013: meeting in Cork, Ireland
- February 2014: meeting in Vienna, Austria
- October 2014: meeting in Zagreb, Croatia
- February 2015: meeting in Brussels, Belgium
- October 2015: meeting in Berlin, Germany
- February 2016: meeting in Amsterdam, Netherlands
- October 2016: meeting in Rome, Italy

Workshop

- iFAAM workshop on risk assessment (Brussels 2015) : "Probabilistic Risk Assessment for Food Allergy: improvement and prospective in the iFAAM project"

Posters

Conference

- **Birot, S.**, Madsen C.B., Christensen T., Crépet A., Kruizinga A.G., Brockhoff, P.B. *Using R for allergy risk assessment in food product.* UseR!: the R User Conference 2015, June 30 - July 3, 2015, Aalborg, Denmark http://user2015.math.aau.dk/poster_session

iFAAM poster competition

- **Birot, S.**, Madsen C.B., Christensen T., Crépet A., Kruizinga A.G., Brockhoff, P.B. *Using R for allergy risk assessment in food product.*
- **Birot, S.**, Madsen C.B., Christensen T., Crépet A., Kruizinga A.G., Brockhoff, P.B. *Food consumption data for use in food allergy risk assessment: a statistically based method for grouping data within and between countries.*

The following paper was prepared in collaboration with partners from other Work Packages during the Ph.D. period. This paper is not part of the methodological developments included in this project and will not be further addressed:

- Pyrz K.M., Austin M., **Birot S.**, Bloom M., Hourihane J.O'B., Kruizinga A.G., Marco Martin G., Mills C., Ramos C., Regent L., Remington B., Turner P.J., Dunn-Galvin A. *How can we better integrate Sex and Gender aspects into Food Allergy Investigation, Assessment and Management? A trans-disciplinary review.*(to be submitted)

Acknowledgements

I would like to thank my supervisor Per B. Brockhoff for all the help, support and advice during all these years. You also have been a great supervisor not only regarding work but also humanly.

I would also like to thank my co-supervisor Charlotte B. Madsen for all support and the advices you have gave me through out the three years.

I would also thank the Work Package partners for the fruitful collaboration and the interesting discussions and questions during the iFAAM meetings.

Finally, I would like to thank my colleagues, friends and family for all the support in my everyday life and to make my life as much enjoyable and interesting as possible.

Contents

Chapter 1

Introduction

1.1 iFAAM project

A short summary of the iFAAM project is disclosed in order to understand the context in which the work during the Ph.D. project was carried out:

"Up to 20 million European citizens suffer from a food allergy. Management of the condition is difficult because of the lack of evidence as to how to either prevent a food allergy from developing, or adequately protect those who are already allergic. The iFAAM project seeks to address this problem by developing evidence based management strategies through a multidisciplinary approach. The project builds on an earlier research study, EuroPrevall. Together these projects are the biggest study of food allergy in the world, involving the world's leading experts in Europe, the USA, and Australia. It comprises over 38 organisations, including patient groups, healthcare professionals and clinicians, risk managers and assessors and the food industry.
The main objectives of the iFAAM project are:

1. *Develop evidence-based approaches and tools for the management of allergens in food.*

2. *Integrate knowledge derived from their application into food allergy management plans and new health advice on nutrition for pregnant women, babies and allergy sufferers.*
3. *Develop strategies to reduce the burden of food allergies in Europe."*

http://research.bmh.manchester.ac.uk/iFAAM/about/iFAAMbrochure.pdf

1.2 Aims of the thesis

The PhD project is part of the Work Package "Risks models" in the module "Modelling allergenic risks". The overall goal of the work package is to develop a validated tiered risk assessment and evidence-based risk management approach for food allergens in the food chain. This work was performed in close collaboration with several project, especially with TNO (Netherlands) and ANSES (France). That's why, this work is closely connected to the work made by other organizations within the Work Package. A summary of some work done by some work packages partners is made in the introduction as it gives inside to some understandings about the context of the investigations and about the choices made in the thesis.

Specifically, the first part of the project focuses on how to utilize consumption data from different countries with the aim of being able to come up with risks evaluations that may work cross nationally. Some work with harmonization of available consumption data is prerequisite to performing cross countries risk assessment. Grouping the food types was one of the way to be able to evaluate the risk across countries. Moreover, the tiered risk assessment approach has two levels (figure 1.1) for which the work on grouping food items was helpful.

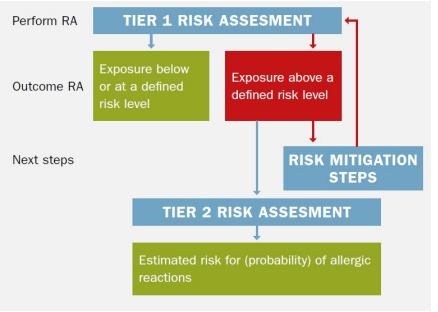


Figure 1.1: Approach in tiered risk assessment (from Marty Bloom, TNO)

The first tier (TIER 1) is a simplify approach that requires minimal input from the users and that should use as much as possible as the available data. Thus, a single point estimate is used as input for the consumption for each food group. Grouping food items reduces the choices to 62, so selecting the correct food types is made easier. Moreover, the food groups allows a cross national risk assessment. The TIER 1 has two possible outcome: green the situation is safe and a red alert is an indication for further step.

In case the exposure is above a defined level of risk, a TIER 2 risk assessment can estimates the probability of allergic reaction for the contaminated product. Risk management actions can be made on the basis of the risk estimation, such as product withdrawal or different product labelling. The TIER 2 requires more inputs form the user and use the whole distribution to estimate the risk of allergic reaction. The TIER 2 is usually performed within organizations that have expertise in allergen risk assessment and which give advice to food producers or public authorities. The work made on grouping and harmonizing food consumption data across countries allows to estimate the risk of allergic reaction for several countries at the same time.

In the second part, the focus was made on comparing and analysing the different approaches for the probabilistic risk assessment (TIER 2). The allergenic risk is modelled combining three sources of information:

- the challenge data: what is the risk of getting an allergic reaction if challenged?
- the contamination data: what level of allergen contaminations are found in various types of food?
- the consumption data: what amount of various types of foods are being consumed (the food groups for categorizing the data from the different countries are used to evaluate the distribution of the different type of food)

Different hierarchical probabilistic methods were compared. The recommended approach was implemented in the open source software and shared with the Work Package partners.

1.3 Outline of the thesis

The thesis consists of 6 chapters that provide an overview of the work completed during the Ph.D. studies. This work was done within the food allergies field. The issue faced within this field with the "may contain" labelling is explained and the connexion with the risk assessment is introduced in chapter 2. A significant part of this work concerns the development of an automated procedure to create food groups suitable for allergen risk assessment, some methodological points and the two submitted papers resulting from the investigations are presented in chapter 3. In chapter 4, the two main methods that have been used for the probabilistic risk assessment are compared, a different approach is proposed and the most suitable method is recommended. This study is presented in a paper that will be submitted. Chapter 5 presents a Shiny application resulting from the work made in chapters 3 and 4. This chapter is presented as a tutorial,so the probabilistic risk assessment can be performed by project partners with no statistical and software knowledge. Concluding remarks are given in chapter 6. Some investigations not included in the thesis and the R code for the shiny application are listed in appendices A- C.

Chapter 2

"May contain" labelling and allergen risk assessment

Around 3-5% of adults and 8% of children worldwide suffer from food allergy [7]. Therefore, food allergy is a significant public health issue and risk management must be conducted in order to limit food hazards for allergenic consumers.

2.1 Food allergy cause and treatment

Food allergy is an adverse reaction where an immune mechanism is involved. The reaction is induced by an immune-mediated sensitivity to food protein. There are 2 types of immunological response, there can be divided into IgE mediated and non-IgE mediated. As the IgE mediated reactions result in a immediate reaction after food intake and may elicit severe reaction (until anaphylaxis), these reaction are the greatest public health concern [10]. However, the non-IgE mediated reactions to food ingestion mostly induce non-life-threatening gastrointestinal symptoms. The risk when accidentally exposed is higher in IgE mediated allergy as it can lead to more severe consequences.

2.1.1 Allergenic reaction process and foods involved

The development of an IgE mediated food allergy is a two steps process. The first is sensitisation and leads to the production of IgE antibodies specific to one or more proteins in a food. The second step is elicitation and occurs when a previously sensitized individual is re-exposed to the same food or food proteins. An allergic reaction occurs when IgE antibodies to the target proteins triggers a cascade of events that leads to the signs and symptoms. Any food can induce an allergenic response and more than 150 different foods have been identified. However, the majority of reactions are elicited by a small number of food items. The following 8 food groups are the most common causes of allergy worldwide: milk, egg, peanut, tree nuts, wheat, soy, fish and shellfish [10]. These allergens may differ as the prevalence of individual foods may differ from age and geographical origin.

2.1.2 Symptoms of food allergy

Food allergies may present clinically with a range of symptoms that can occur either in an isolated form or associated. A large variety of symptoms may be developed and they can involve the digestive, respiratory, cardiovascular or cutaneous organ systems. Allergenic reaction severity from mild to severe: Oral Allergy Syndrome (OAS), gastrointestinal symptoms, skin symptoms, respiratory symptoms and anaphylaxis shock. Reaction severity depends on the sensitivity of the individual, the level of exposure and modifying factors such as the presence of others conditions [10]. We can list several factors influencing the occurrence and the severity of symptoms, such as the physico-chemical characteristics of the allergen, the way the food is eaten, severe or uncontrolled asthma or the level of food specific IgE to the food in question.

2.1.3 Diagnosis of food allergy

A careful family and clinical history are the basis for diagnosis of food allergy. Food dairies, skin prick tests (SPTs), allergen specific IgE measurements, food elimination diets and food challenges are the part of the standard protocol for the diagnosis of food allergy. A positive SPT indicates sensitisation to the tested food, but it is not diagnostic for food allergy. Allergen-specific serum IgE antibodies indicate sensitisation to a particular food, but are not diagnostic without a clinical history or food challenge [10]. Diagnosis is confirmed by exclusion of the suspected food and the subsequent amelioration of symptoms,

and by the reappearance of symptoms on re-introduction of the offending food, ideally in double-blind placebo controlled food challenges (provided that the initial symptoms were not life threatening).

2.1.4 Food allergy management

Currently there is no cure or effective treatment for food allergies, the reference treatment is avoidance of food containing the allergen after identification of offending foods. Successful avoidance depends on the public having complete and accurate information on the substances in a food, information obtained from food labels [10]. Rescue medication to treat an eventual reaction due to accidental food ingestion may occur as avoiding food allergens can be a challenge. Immunotherapeutic approaches are currently under development and a curative may be successfully developed in the future.

2.2 “May contain” labelling description and usage for industry

The protection of allergenic consumers stands by helping the consumer identifying foods containing food allergens. As we have seen previously that allergen avoidance is the main remedy against allergenic reaction.

2.2.1 “May contain” labelling features

World-wide regulatory initiatives have been aimed at mandatory declaration of the most important food allergens. So thanks to an accurate and unambiguous labelling of food products, allergenic consumers are able to identify correctly foods containing allergens. Within the European Union, all ingredients in a food should be in the list of ingredients. A directive specifies further rules of labelling of the 11 most common allergenic foods: cereals containing gluten, crustaceans, eggs, fish, peanuts, soy beans, milk, tree nuts, celery, mustard, and sesame and ingredients derived from those foods [16]. However this legislation only concerns known ingredients. Unfortunately the presences of allergens in food products resulting from unintended contamination can also threaten the health of allergenic consumers. Thus, food safety directive and regulation require that foodstuff containing allergenic ingredients not indicated on the label are unsafe for consumers with a food allergy and should not be placed on the

market. However, advisory statements are voluntary, in contrast to ingredient listing. Although uniform wording of advisory warnings is recommended, the guidelines are voluntary and have done little to reduce the prevalence and variety of advisory labels currently used. Examples of advisory labels are displayed in table 2.1 [16].

Examples of advisory warnings found on food labels
May contain ...
May contain traces of ...
Produced in a factory which handles ...
Produced on shared equipment which also processes ...
Made in a production area that also uses ...
Made in a factory that also produces ...
Not suitable for ... allergy sufferers
Packed in an environment where ... may be present
Due to the methods used in the manufacture of this product, it may occasionally contain ...

Table 2.1: Examples of advisory warnings

2.2.2 "May contain" labelling usage for industry

The concept of managing food allergens as a food safety risk emerged in the last decade of the 20th century and has matured considerably over the last 10-15 years. At the beginning, few things were known about the key determinants of risk: how sensitivity and reactivity varied across the allergenic population in response to the dose consumed. Industry's approach to date has been based around existing Good manufacturing Practices assuring segregation of allergenic ingredients and systematic declaration of allergens on labels when mandated. To measure the possibility of cross-contamination, the UK Food Standard Authority produces a comprehensive guide to best practice that recommends a non-quantitative approach to determining whether there is a possibility of cross-contamination and if so, to provide an advisory warning [17]. The standard control of allergen cross-contamination is based on visual inspection of the production line and of the final product. In the absence of knowledge about the levels of allergens required to provoke adverse reactions, many manufacturers have adopted an ostensibly "fail-safe" approach using precautionary labelling. Initially welcomed as helpful by allergic consumers, the increasing and inconsistent use of this type of warning across product types and sectors has considerably diminished its impact as a risk reduction tool [17]. This has led to consumers being increasingly frustrated with precautionary labelling and taking risks.

2.2.3 Allergen controls in practice

While accurate labelling may seem a straightforward issue, incorrect labels actually account for a large proportion of product recalls in the UK, as well as in others countries. The two main causes of recalls are the omission of an allergen on an ingredient label and placing the wrong product in the wrong pack [17]. The number of allergen incidents range from to 129 (2012) to 79 (2010), although few resulted from a reaction in a consumer. The breakdown of allergen incidents by allergen type is shown on figure 2.1 [1]. While milk is the major type of allergen during the last 4 years, 2013 has seen a huge increased in the incident in the peanut category.

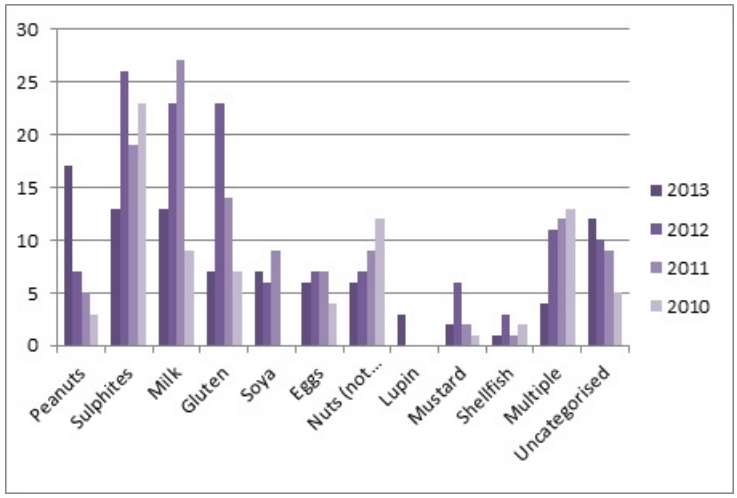


Figure 2.1: Number of allergen incidents reported to the FSA in UK from 2010 to 2013

In order to manage allergens effectively and provide the necessary information to consumers, it is important to adopt an integrated approach to ensure that allergen information is transmitted accurately across the supply chain.

2.3 "May contain" labelling – Consumer perspective

Consumers face complex food choices on a daily basis. Most consumers balance a number of issues when considering which foods to eat, including the price, taste and whether the food is nutritious [4]. Food allergic individuals additionally need to avoid allergens to prevent potentially life threatening reactions. Packet information is vital in assisting with such decisions.

2.3.1 Current state of "may contain" labelling

To help nut allergic consumers avoid products that contain nuts there are several sources of information on food packaging. These include the product name, ingredients list, allergy advice and precautionary information. If there is a possibility that a food may contain traces of an allergenic food, not as an intentional ingredient, but as a result of cross-contamination through, for example, shared manufacturing equipment. This risk is often indicated by a precautionary "may contain" type label (e.g. "may contain traces of peanut"). Such precautionary statements are not regulated either in food safety or food labelling legislation[4]. However there is a general requirement for labelling not to be misleading, and to be safe under general food law.

2.3.2 Consumers critical about "may contain" labelling

At the beginning, "may contain" precautionary labelling was considered responsible and helpful to allergic consumers, but as the range and the number of products labelled increased, consumers began to questions whether the food suppliers were using it defensively "to cover their backs" [3].

The use of precautionary labelling on apparently unconnected products such as pre-packed salads, tomato sauces and many other items, has led to decrease the impact of its message. Discussions with allergic consumers improve the understanding of their consumer behaviour and their allergen avoidance strategies. Many openly declare that they disregard allergen trace contamination information on products. They cannot believe that such a huge proportion of food products on sale may put their lives at risk. A recent study of allergic individuals reported that 8% of those with accidental reactions attributed it to having ignored a precautionary label [15].

Additionally, foods that carry precautionary labels may actually be safe to eat and therefore consumers may be following unnecessarily restrictive diets by observing the warning labels. It is likely 90% of products with cautionary labels will contain no residues of peanut protein, and some of those that do are at levels unlikely to cause a clinical reaction [4]. Thus, allergic consumers can interpret “may contain” labelling as being weaker than the actual meaning and previous experience of a product is an important arbiter of how uncertainty introduced by “may contain” labelling is interpreted. Previous experience is trusted to ensure future safety [4].

Finally, as seen previously there is variation in wording (Table 2.1) induced by the non-regulation of “may contain” labelling. Although this variation is not intended to convey different degree of risk, it is often interpreted by the consumer as doing so [4]. Consumer decisions about whether to trust a product depend on how ingredients and allergen information are presented. Current labelling practices created difficulties for all consumers.

2.3.3 “May contain” labelling improvement

Allergic consumers find difficulties in trusting precautionary labelling and finding food without such labelling [4]. A standardized and clear ingredients labelling is essential for all product. This will help all consumers and will improve the ability of allergic consumers to carry out their own risk assessment and make informed food choices. A standardized labelling is highly valuable as it will help allergic consumers to identify safe products. For example, allergy boxes in a particular size or colour with the information presented in the same order would be beneficial. A “nut-free” label would also be valuable for allergic consumers, as a visible symbol, a general warning prompt to seek out further information from elsewhere on the packet [3].

However, these improvements cannot be made without the support of industries and manufacturers. As the awareness of food allergies is a public health issues is growing, manufacturers are starting to seek to clean up by removing allergens from their production.

2.4 Risk assessment contribution to "May contain" labelling

There is an emerging consensus that, as with other risks in society, zero risk for food-allergic people is not a realistic or attainable option [6] [9]. Food allergy challenge data and new risk assessment methods offer the opportunity to develop quantitative limits for the unintended presence of allergens which can be used in risk- based approaches.

2.4.1 Need for quantitative risk assessment

Attempts to standardise the contents of advisory labels using voluntary guidance have not been successful. The listing of all potential allergens as an ingredient when cross contamination is likely might reduce ambiguity but would further restrict consumer choice. The best solution would be to quantify the hazard: whether the degree of contamination is sufficient to trigger an allergic reaction, and communicate clearly to the consumer [16].

However, a prerequisite is defining a tolerable level of risk on which to base the thresholds to be used for management of allergen risks. For chemicals where the toxicological effect has a threshold, the "acceptable" or "tolerable" daily intake is typically used. With this method, quantification of the risk is not possible. Thus, to reduce the magnitude of the risk and the use of wording such as "considerable" and "appreciable", the risk of allergic reaction fist needs to be quantified using probabilistic risk assessment methods [10]. Based on those assessments and involving all relevant stakeholders, a tolerable risk can be defined.

2.4.2 Evaluating allergy prevalence in population

Challenge data form the basis for risk assessment, but optimal approaches to using these data need to be defined. A key consideration is that those data are generated using human beings. This has an advantage that no extrapolation from other species or test systems is necessary, but also means that certain type of data, such as individual dose-response data, can only be generated to a very limited extent. Modelling approaches make better use of the limited data, for instance by taking into account the whole dose distribution. However the success of such approaches is predicated on developing a better understanding of

how the predictions compare with reality.

Thus, a true safe threshold for any food and person is impossible to determine. However, the aim of quantitative risk assessment is to reduce the risk of harm from cross-contamination to a level considered tolerable, rather than to eliminate the risk altogether [16]. Researchers have therefore developed the concept of Eliciting Dose (ED) combining data from DBPC clinical studies. The the individuals NOAEL (No Observable Adverse Effect Level) and Lowest Observed Adverse Effect Level (LOAEL) data are used to estimate the EDs for each allergen when the data are sufficient. ED01 is the threshold below which less than 1% of the people will react. Only a few studies have been published assessing EDs for the more common food allergens by double blind, placebo controlled food challenged. Nonetheless, for many common food allergens, there is a high degree of agreement between published ED's, the current EDs for the most common allergen are presented in table 2.2 [10]. Further data are expected from high quality double-blind placebo-controlled studies particularly the Euro-PREVALL collaboration.

Allergen	Reference Dose (mg Protein)	Basis of Reference Dose	Quality of Database
Peanut	0.2	ED01	Excellent
Milk	0.1	ED01	Excellent
Egg	0.03	ED01 and ED05 95% lci*	Excellent
Hazelnut	0.1	ED01 and ED05 95% lci*	Good
Soy	1	ED05 95% lci	Sufficient
Wheat	1	ED05 95% lci	Sufficient
Cashew	2	ED05 95% lci	Sufficient
Mustard	0.05	ED05 95% lci	Sufficient
Lupin	4	ED05 95% lci	Sufficient
Sesame	0.2	ED05 95% lci	Marginally sufficient
Shrimp	10	ED05 95% lci	Marginally sufficient

** Lower confidence interval*

Table 2.2: Summary of VITAL Scientific Expert Panel Recommendations

2.4.3 Quantitative risk assessment methods

A risk based approach focuses on the probability that exposure to a food will result in an adverse effect. It should be able to provide the scientific foundation to determine agreed threshold levels for industry to use in risk-management decisions, including the use of advisory warning of cross-contamination. It should

also be capable of being used to decide what level of cross-contamination is high enough to justify a product recall/withdrawal and what level of the allergen is sufficiently low to justify a "free-from" claim [9].

The characteristic of this method is that it uses distributions. The input variables are the 'allergen exposure distribution' and the 'dose distribution curve' from challenge data. The 'allergen exposure distribution' is the combination of the 'distribution of the amount of food consumed' and the 'distribution of concentration of allergen' in the food product under investigation [10].

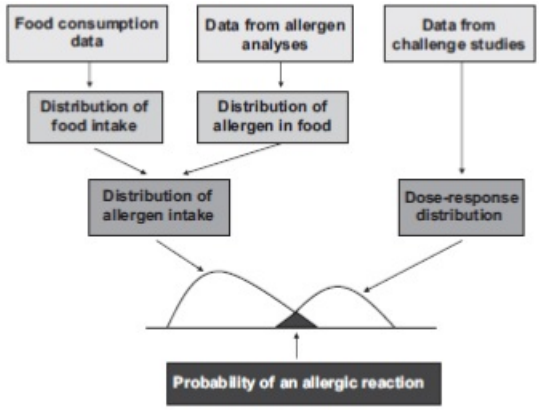


Figure 2.2: Concept of probabilistic risk assessment in food products [10]

The outcome of the probabilistic risk assessment is the probability of an allergic reaction occurring upon consumption of the food product in question. The probability is a numerical value that estimates the magnitude of the risk. The advantage of this method is that it makes the basis of the value judgement that the risk manager must make very explicit if this specific risk is acceptable or not, or what concentration is acceptable.

Probabilistic modelling is considered to be the most promising approach for use in population risk assessment [9]. For all approaches, further improvement of input data is desirable, particularity data on consumption patterns/food choices in food-allergic consumers, data on minimum eliciting doses and data that can be used to evaluate whether the whole population at risk has been modelled accurately.

2.4.4 VITAL: an example of a risk assessment tool to improve labelling

Australia and New-Zealand have already started to use EDs values to improve labelling. In 2007, the food manufacturing industry, with the input of consumer groups and regulatory authorities, developed a standardized risk assessment tool called Voluntary Incidental trace Allergen Labelling (VITAL). This allows manufacturers to assess potential cross-contamination quantitatively and determine the need for advisory warnings [16].

Thresholds levels were based on published LOAEL data. When the amount of allergen of allergen present is above the threshold level, but not at sufficient amounts to be listed as an ingredient, manufacturers use an advisory statement with the format "may be present". No advisory warning is recommended if levels are lower than this cut-off. Although some very sensitive people might react to levels of allergen below the threshold, these people are in general more likely to avoid potentially problematic foods. The scheme means that advisory warnings are used only when warranted and that the warnings are standardised, so providing clear and simple information to consumers.

Studies from Europe and US assessing cross-contamination have found that most foods with advisory warnings for allergens not listed as an ingredient do not contain allergen levels above VITAL thresholds. Thus, adoption of the VITAL scheme in Europe would mean that most products with advisory labels would no longer require them [16].

Chapter 3

Food groups for allergen risk assessment

3.1 Aim and outline of the chapter

As explained in chapter 1, food groups are needed as an input for the TIER 1 risk assessment. The TIER 1 risk assessment is a simplify version of the risk assessment: the exposure is calculated with a consumption point estimate and a concentration point, which is then compared to a single threshold dose of allergen (selected by food allergen experts).

The number of food items can be very high and very details. Thus, groups of food items are created to ease the user choice. Furthermore, the creation of a pan-European database is investigated, as there were no previous studies assessing the feasibility of merging National Food Consumption Surveys in an automatic way for allergen risk assessment. In the allergen risk assessment fields, the focus was actually made on identifying what could be improved for the threshold distribution. The food groups can also be used in the probabilistic risk assessment if a risk for a food group needs to be calculated.

Using the National Food Consumption Surveys from Netherlands, France and Denmark, an approach is created for answering the different aims of this investigation. In this chapter, the different investigations conducted to elaborate this approach are presented with two journal articles:

- "Grouping food consumption data for use in food allergen risk assessment" submitted in Journal of Food Composition and Analysis. The methodological details of the approach are detailed in this article. Food groups created for Netherlands, France and Denmark illustrate the steps and the outcome of the procedure.

- "Combining food consumption data from different countries for creating food groups for allergen risk assessment (in Europe)" submitted in the Nutrients journal . In this paper, the procedure explained in the first paper is applied to the food consumption of the three country combined. Some methodological investigations on merging National Food Consumption surveys with different designed are addressed.

To explicit the process of the choices, the investigations on the decision criterion and the number of days in the surveys not detailed in the two articles are presented in this chapter as additional information to the two articles.

3.2 Presentation of the National Food Consumption Surveys

National Food Consumption Surveys from Netherlands, France and Denmark were used in this project. National Food Consumption Surveys do not indicate which individuals are allergic to which allergen and food consumption for allergic individuals are not recorded in specific surveys. That's why, the assumption is made that allergic individuals have the same consumption patterns as the non allergic ones recorded in the regular surveys in the three countries. Thus, these surveys are used to estimate the consumption of allergic individuals.

European guidelines are used to record the consumption of the different foods. Especially, the same coding system was used to code the foods in the three different countries. Thus, each recorded food is matched to the corresponding food item in the FoodEx 2 coding system. So, merging the three consumption databases is easier, technically. The processed food items were selected as cross contamination is not expected in raw food such as raw vegetables or fresh fruits. Furthermore, some of the food items doesn't have enough occurrences, so food items with less than 5 occurrences were merge to similar food items within the same initial TNO group. In the table 3.1, the number of participants and the number of food items in each original survey are indicated.

Country	Numbers of participants	Numbers of survey food items
Netherlands	3819	618
France	2624	391
Denmark	2029	227
All countries	8472	813

Table 3.1: Number of participants and food items in each survey

Unfortunately, the surveys are not designed in the same way in all countries. Netherlands perform 2 non consecutive 24 hour recalls, whereas a pre-coded 7 days food record is done in France and Denmark. For each consumer, the maximum consumption on a single eating occasion is used. It is well known that the maximum over 7 days is higher than the maximum over 2 days, as there is more chances to notice higher consumption over 7 days than over 2 days. Thus, it was also checked in this chapter as a preamble to the chapter that the survey design has a limited impact on the risk assessment outcome.

3.3 Procedure's steps for grouping food items

The procedure for creating groups of food items is reminded shortly in this section (figure 3.1), further methodological details can be found in the article 1.

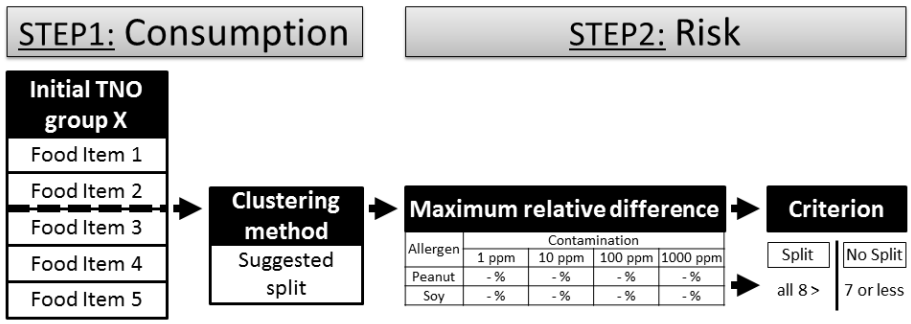


Figure 3.1: Procedure steps – summary (from article 1)

3.3.1 Step 1: clustering food items

TNO (one of the work package partner) designed food groups for allergen risk assessment using a combination of expert knowledge and summary statistics on food items consumption. These groups were used as a basis to the automated procedure for grouping the food items in the National Food Consumption Surveys in the three countries.

During the clustering step (first step), a division of the initial group is suggested using a customized clustering method based on consumption only. The similarities between consumption distributions are evaluated with two statistical tests that are commonly used to check if two distributions are drawn from the same distributions: the Kolmogorov-Smirnov test [5] and the Cramér-von Mises test [2]. In short, we are creating groups of food items that have homogeneous consumption patterns and that are also meaningful.

3.3.2 Step 2: risk simulation and decision

In the second step, the risk of allergic reaction is calculated for each initial group and its subgroups defined by the clustering method. We will then check if the subgroups based on the consumption, only, are also homogeneous in term of risk.

As described in chapter 1, the existing statistical methods for a validated tiered risk assessment all have 3 different data inputs:

- Challenge data (threshold distribution in the population for each allergen)
- Contamination data (how much allergen is in the contaminated product)
- Consumption data (the amount of contaminated product which is consumed among the population)

From these 3 inputs, the percentage of the population likely to have an allergenic reaction is estimated, i.e. what is the chance for an allergenic consumer to eat a high enough amount of allergen to trigger a reaction. The inputs distribution that are actually used to calculate the risk of allergic reaction are described in the articles 1 and 2. A fixed range of concentration levels and the same allergen distributions are used in a framework, thus only the consumption changes and the impact of the differences in consumption on the risk can be evaluated. Furthermore, the way the simulations are performed are detailed in article 1. During the investigation two distributions were used to fit the threshold data: the log-normal distribution and the Weibull distribution. However, only the

Weibull distribution was used in articles 1 and 2, the reasons are explained in the following section.

Finally, a criterion is designed to assess the differences in risks across food groups or countries. And, decide if an initial group is homogeneous enough to estimate the risk of allergic reaction or if the group should be divided. Some investigations on how the criterion was selected are presented in this chapter.

3.4 Assessing differences in risk of allergic reaction across food groups with decision criteria

The investigations made to decide on the decision criterion are presented in this section. The investigation performed in article 2 is used to illustrate the choice, In article 1, similar investigations are performed for the within country food group outline.

3.4.1 Groups checking aim and method

Groups checking aim

In this part, we will check if the clusters based on only the consumption also have homogeneous risks. It should be checked if the clustering aims are all fulfilled:

- if the subdivision of an initial TNO group is needed, i.e., the risk is different for the subgroups of one TNO group or a risk for the overall TNO group is representative enough
- and within one initial TNO group, if the risk is the same for all the countries and if common food groups make sense. I.e, if the same grouping can be applied to all countries or if there are some differences between countries.

Group checking method

The aim of the food item groups is to have food groups with homogeneous consumption patterns and also with similar risk outcome:

- The 1st step will be to investigate the consequences of splitting on the consumption survey of combined countries, as the aim is to build a pan-European food consumption database.
- Then the consequences of the group outline on the combined countries consumption will be investigated on the risk at a country level. And finally, it will be checked if for a TNO group all countries have similar risks

We can directly use the risk calculated for each subgroup and compare it to the overall risk of the initial TNO group.

3.4.2 Relative and absolute difference calculation

In order to asses the homogeneity of risk within each TNO group, we will calculate the maximum relative and absolute difference between each subgroup and the risk of each TNO group.

Maximum relative difference

$$relative = max(\frac{|Risk_{Subgroup} - Risk_{Group}|}{Risk_{Group}}) \tag{3.1}$$

Maximum absolute difference

$$absolute = max(|Risk_{Subgroup} - Risk_{Group}|) \tag{3.2}$$

3.4.3 Summary statistics across food groups

The risks are calculated for each groups and their subgroups defined with the clustering method. However, we need to evaluate whether the difference between the group risks and each subgroup's risk is high, so splitting an initial group is relevant. Unfortunately, there are no references to assess the size of the differences between risks of allergic reaction for a food groups and its subgroups. That's why, some summary statistics on maximum relative and absolute difference are displayed in order to assess the trends and help to choose a decision criterion for whether a group should be split or not. The maximum,

95% quantile and the median of those differences are selected to picture their distributions.

Maximum

Distribution	Allergen	1	10	100	1000
LogNorm	Peanut	642.14	567.15	419.43	248.81
LogNorm	Soy	807.36	654.68	644.08	559.31
Weibull	Peanut	226.61	221.49	208.10	178.32
Weibull	Soy	235.92	236.76	232.62	222.94

Table 3.2: Maximum relative difference across food groups

Distribution	Allergen	1	10	100	1000
LogNorm	Peanut	2.47	8.30	19.81	36.91
LogNorm	Soy	0.08	0.77	3.63	11.75
Weibull	Peanut	3.34	7.65	16.48	30.75
Weibull	Soy	0.81	1.96	4.65	10.58

Table 3.3: Maximum absolute difference across food groups

95 quantile

Distribution	Allergen	1	10	100	1000
LogNorm	Peanut	112.58	75.80	54.29	37.31
LogNorm	Soy	227.05	157.89	110.01	71.52
Weibull	Peanut	48.98	47.77	44.69	37.75
Weibull	Soy	50.44	49.98	49.18	47.34

Table 3.4: 95 quantile relative difference across food groups

Distribution	Allergen	1	10	100	1000
LogNorm	Peanut	1.97	6.84	14.11	19.94
LogNorm	Soy	0.04	0.48	2.98	9.58
Weibull	Peanut	2.59	5.72	11.18	18.21
Weibull	Soy	0.65	1.57	3.62	7.74

Table 3.5: 95 quantile absolute difference across food groups

Median

Distribution	Allergen	1	10	100	1000
LogNorm	Peanut	27.96	21.02	14.20	8.92
LogNorm	Soy	42.85	32.75	27.00	20.07
Weibull	Peanut	11.66	11.31	10.62	8.90
Weibull	Soy	12.35	12.21	11.85	11.41

Table 3.6: Median relative difference across food groups

Distribution	Allergen	1	10	100	1000
LogNorm	Peanut	0.49	1.97	4.43	7.05
LogNorm	Soy	0.01	0.10	0.75	2.74
Weibull	Peanut	0.75	1.68	3.45	6.37
Weibull	Soy	0.19	0.45	1.03	2.33

Table 3.7: Median absolute difference across food groups

1% risk criteria

In order to check if all the groups have an high enough risk to be split. We will check the number of subgroups in each TNO group with a risk lower than 1%. If all the subgroups in the group have a risk lower than 1%, then it is not needed to split further, as the risk would be low anyway, they would be no impact of dividing an initial group.

Distribution	Allergen	1	10	100	1000
Log-Normal	Peanut	15	1	0	0
Log-Normal	Soy	34	31	10	1
Weibull	Peanut	0	0	0	0
Weibull	Soy	15	1	0	0

Table 3.8: Number of TNO groups with all subgroups with risk lower than 1 percent across food groups

3.4.4 Decision criteria

Observations

In order to decide whether a group need to be split, a decision criteria is necessary. It will help to assess if the differences between the overall group's risk and the subgroups' risks are high.
Different observations can be made on the summary (across initial TNO groups) statistics presented above:

- **Maximum relative difference:** the Log-Normal and Weibull distribution give different outcomes. It is known that the low contamination levels are less detected by the Log-Normal distribution than the Weibull distribution (ref Rimbaud). Thus, the maximum relative difference is very high for low contamination and for the Log-Normal distribution. And, more stable when the contamination is increasing for the Log-Normal distribution, and always stable for the Weibull distribution. And, as expected the maximum relative difference is more sensitive with low contamination and Soy allergen.
- **Maximum absolute difference:** for the low contamination, the maximum absolute difference is really dependent on the distribution, the allergen and the contamination. The difference increase with contamination. And the difference is higher for Peanut allergen than for Soy Allergen, as the peanut threshold distribution is more sensitive to low contamination.
- **1% risk criterion:** there are few TNO groups with risk lower than 1%, most of them when the contamination is low. So there are no food groups with a very low consumption that does not need to be split. As the risk could be very low, it wouldn't matter if groups are split or joined.

The observations can be made for the 3 different summary statistics presented above (Maximum, 95 quantile and Median).

Decision criteria choice

For further decisions, we will only use the Weibull distribution and the relative difference to make a decision. As it has been noticed, all the summary statistics are more stable across all the conditions introduced to evaluate the impact of clustering on the risk assessment, only for the Weibull distribution

and the maximum relative difference between the group's risk and subgroups' risks. Thus, in order to have a unique value as the threshold, the Weibull median of the maximum relative difference between risks on the combined countries is calulated and used as the threshold. Thus, the actual value to decide whether a group should be split is around 12.6%. Six groups with a low number of food items, with no subgroups suggested, are removed from the median calculation.

This number can be interpreted as such: if the relative difference between the overall group risk with the subgroups risks is higher than the threshold (around 13%) then the group is split. In order to assess the number of groups that need to be split. For each TNO group, this threshold will be compared to the maximum relative difference between risks calculated for each contamination levels, only for the Weibull distribution. If the risk of all the 4 contamination levels and the 2 allergens are above the threshold, then a subgroup is decided to be necessary.
To illustrate the method, we can use the peanut butter example:

Allergen	1	10	100	1000
Peanut	33.8	32.8	30.1	24.5
Soy	35.4	34.9	34.3	32.5

Table 3.9: Maximum relative difference for the Peanut butter, nut paste group (Weibull distribution)

In this case, the maximum difference for all the contamination levels and the 2 allergens are higher than the threshold, i.e. (around 12.6%). So this group should be split.

Change in the decision criteria (from median to 100%)

This threshold can be adjusted according to the decision criteria used to decide if a group should be split. A short study of the consequences of changing the decision criteria on the number of subgroups is presented in this section.
In order to investigate the impact of the threshold on the number of group to be split, we will vary the decision critria from 50 to 100%.

Criteria decision	Thresholds value	Number of groups	Percentage of groups
0.50	12.56	18.00	39.13
0.55	16.24	15.00	32.61
0.60	18.59	13.00	28.26
0.65	20.01	11.00	23.91
0.70	22.63	9.00	19.57
0.75	25.97	4.00	8.70
0.80	28.69	3.00	6.52
0.85	30.89	3.00	6.52
0.90	35.39	3.00	6.52
0.95	49.96	1.00	2.17
1.00	236.76	0.00	0.00

Table 3.10: TNO groups with subgroup needed when the threshold is varying

3.5 Surveys designs comparison

Article 2 refers to some investigations made to assess the impact of the number of days in the consumption surveys on the risk estimation. Thus, the investigations to decide from which level the number of days in the consumption surveys have an impact on the consumption distribution are detailed in this section.

The method used in article 2 to compare the surveys design is reminded shortly: in order to compare the surveys design, 2 days are sampled from the French and Danish National Food Consumption Surveys 50 times. Then, the maximum consumption over the days is calculated and averaged over the 50 times. The food items' consumption distribution of the original and re-sampled are then compared in order to assess the impact of the number days on the maximum consumption. More explanation on this investigation can be found in article 2.

3.5.1 Relative difference in median

When we compare the median between original and re-sampled data on figure 3.2, we can see that most of the differences are lower then 5%. Therefore, it seems that the number of days in the survey does not have high impact on the food item distribution.

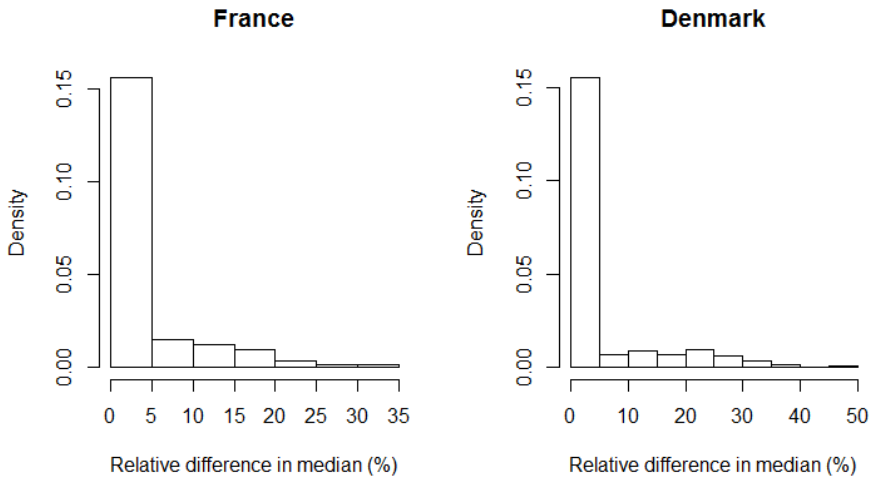


Figure 3.2: Relative difference in median between the original and the resampled food items consumption distributions in France and in Denmark

3.5.2 Chocolate example

The consequences on the risk assessment outcome are investigated first using an example: the chocolate products group. Initially, a different concentration level was used to calculate the risk of allergic reaction. But, as the level of concentration remains constant and the aim is to assess the difference in consumption between the original and the re-sampled consumption distribution, the level of contamination selected to perform this study is not expected to have consequences on the decisions.

The number of consumers, relative difference in median and the risk calculated for the original and re-sampled consumption distributions are presented in table 3.11. None of the food item in the chocolate group have high different consumption distribution or risks when we re-sampled the survey days.

Distribution	Name	Number of consumers	Relative difference in median	Log normal		Weibull	
				Original	Re-sampled	Original	Re-sampled
A034F	Chocolate	288	0%	0.35‰	0.33‰	0.70‰	0.68‰
A034P	White chocolate	21	0%	0.32‰	0.29‰	0.65‰	0.63‰
A034Q	Filled chocolate	501	14.84%	0.33‰	0.30‰	0.67‰	0.64‰
A034R	Chocolate coated confectionery •	25	0%	0.32‰	0.32‰	0.68‰	0.67‰
A034G	Bitter chocolate	85	0%	0.15‰	0.14‰	0.46‰	0.45‰
A034H	Bitter-sweet chocolate	257	0%	0.18‰	0.17‰	0.50‰	0.49‰
A034J	Milk chocolate	587	0%	0.23‰	0.21‰	0.56‰	0.54‰
A034L	Cream chocolate	198	0%	0.74‰	0.73‰	1.08‰	1.09‰

Table 3.11: Mean percentage of allergic reaction for chocolate group for the original and re-sampled data in France

As expected, the risk assessment calculated with the re-sampled survey is slightly lower than the RA done with the 7 days.

3.5.3 Evaluate the impact of differences

We will add noise to the consumption in the chocolate food group in order to evaluate the impact of various range of noise on the risk assessment outcome. We will simulate scenarios from 5% to 35% noise. Note that these scenarios could be more conservative as the actual 2 days re-sampling, as the whole distribution is shifted from 5% up to 35%.

	Original data	5% noise	10% noise	15% noise	20% noise	25% noise	30% noise	35% noise
A034F	0.35‰	0.34‰	0.34‰	0.31‰	0.30‰	0.29‰	0.27‰	0.26‰
A034P	0.32‰	0.30‰	0.29‰	0.28‰	0.27‰	0.26‰	0.25‰	0.24‰
A034Q	0.33‰	0.31‰	0.30‰	0.29‰	0.28‰	0.26‰	0.25‰	0.24‰
A034R	0.32‰	0.31‰	0.30‰	0.29‰	0.28‰	0.26‰	0.25‰	0.24‰
A034G	0.15‰	0.15‰	0.14‰	0.13‰	0.13‰	0.12‰	0.12‰	0.11‰
A034H	0.18‰	0.18‰	0.17‰	0.16‰	0.15‰	0.15‰	0.14‰	0.13‰
A034J	0.23‰	0.22‰	0.21‰	0.21‰	0.20‰	0.19‰	0.18‰	0.17‰
A034L	0.74‰	0.72‰	0.68‰	0.67‰	0.65‰	0.62‰	0.60‰	0.56‰

Table 3.12: Risk of allergic reaction calculated for the different of noise added to the consumption distribution with the Log-Normal distribution used to fit the threshold data

	Original data	5% noise	10% noise	15% noise	20% noise	25% noise	30% noise	35% noise
A034F	0.70‰	0.70‰	0.68‰	0.66‰	0.65‰	0.63‰	0.61‰	0.60‰
A034P	0.65‰	0.64‰	0.63‰	0.62‰	0.61‰	0.59‰	0.58‰	0.56‰
A034Q	0.67‰	0.66‰	0.65‰	0.62‰	0.61‰	0.61‰	0.59‰	0.57‰
A034R	0.68‰	0.67‰	0.66‰	0.64‰	0.63‰	0.61‰	0.59‰	0.58‰
A034G	0.46‰	0.45‰	0.45‰	0.44‰	0.43‰	0.41‰	0.41‰	0.40‰
A034H	0.50‰	0.49‰	0.49‰	0.47‰	0.46‰	0.45‰	0.45‰	0.42‰
A034J	0.56‰	0.55‰	0.53‰	0.53‰	0.51‰	0.50‰	0.49‰	0.47‰
A034L	1.08‰	1.06‰	1.05‰	1.02‰	1.01‰	0.97‰	0.95‰	0.45‰

Table 3.13: Risk of allergic reaction calculated for the different of noise added to the consumption distribution with the Weibull distribution used to fit the threshold data

3.5.4 Interpretation and decisions

The noise is simulated from 5% to 35% in order to match the range of difference in median found in the French National food consumption survey. The outcome of the risk calculator is quite close to the one calculated from the re-sampled survey. However from a 25% noise, the risk calculated is starting to be different.

Thus, in the article 2, we will focus on the food item with difference in median higher than 25%. The re-sampling impact on the risk outcome for those food items is assessed by calculating the risk for both original and re-sampled data for those food items. Summary tables for those food items are included in article 2, so it can be checked that the differences in consumption are not high enough to have an impact on the risk outcome, even when the differences between original and re-sampled consumption are high.

3.6 Article 1: Grouping food consumption data for use in food allergen risk assessment

Title: Grouping food consumption data for use in food allergen risk assessment

Authors

Sophie BIROT[a], Charlotte B. Madsen[b], Astrid G Kruizinga[c], Tue Christensen[b], Amélie Crépet[d], Per B. Brockhoff[a]

Authors' affiliations

[a] DTU Compute, Richard Petersens Plads, DK-2800 Kgs. Lyngby, Denmark sobi@dtu.dk

[b] National Food Institute, Technical University of Denmark, Denmark

[c] The Netherlands Organization for Applied Scientific Research (TNO), Zeist, The Netherlands

[d] ANSES, French Agency for Food, Environmental and Occupational Health Safety, 14 rue Pierre et Marie Curie, 94701 Maisons-Alfort, France.

Highlights (max 85 characters incl. spaces)

- Procedure for creating food groups for probabilistic allergen risk assessment
- Customized clustering method followed by allergen risk assessment per food group
- National Food Consumption Surveys from France, Netherlands and Denmark
- Danish consumption data available per food group

Abstract (max 200 words)

Food allergic subjects need to avoid the allergenic food that triggers their allergic. However, foods can also contain unintended allergen. Food manufacturers or authorities need to perform a risk assessment to be able to decide if unintended allergen presence constitutes a risk to food allergic consumers. One of the input parameters in risk assessment is the amount of a given food consumed in a meal. There has been little emphasis on how food consumption data can be used in food allergen risk assessment. The aim of the study was to organize the complex datasets from National Food Consumption Surveys from different countries (France, Netherlands and Denmark) to be manageable in food allergen risk assessment. To do this, a two-step method was developed. First, based on initial groups of similar food items, the homogeneity of consumption

was evaluated using a customized clustering method. Then, the risk was calculated for each initial food group and its subgroups to verify if it also represents a relevant difference in risk. 48 food groups were designed in Denmark (53 in the Netherlands, 54 in France). Finally, summary statistics and names for each food group for the Danish data illustrate the results when applying the procedure.

Key words

Food allergy; National Food Consumption Surveys; food groups; probabilistic risk assessment

1. Introduction

In Europe the lifetime prevalence and the point prevalence of self-reported food allergy is around 17% and around 6%, respectively. These figures probably reflect an overestimation of the true prevalence that also varies between countries, but the true prevalence is not known (Nwaru et al., 2014). In food allergic subjects' exposure to food allergens can trigger an acute allergic reaction, which can have a wide range of symptoms: from itch in the mouth to life threatening anaphylactic shocks (Bock et al., 2001). For this reason allergic consumers need to avoid consumption of the food containing the harmful allergen (Fernández-Rivas and Asero, 2014). This can be done by avoiding foods where the allergen is in the list of ingredients as the European Directive 2003/89/EC makes the labelling of all ingredients mandatory. However contamination during the production may occur and as the allergen is not part of the ingredients, it is not mentioned in the ingredient list. Thus, products with unintended allergen are potentially harmful for the allergic consumers. The increasing use of "may contain" labelling could help allergic consumers manage their food allergy. However, the lack of regulation in Europe has led to a misuse of such labels. The labels are actually used even though many products are not contaminated (DunnGalvin et al., 2015).

To be able to decide if the unintended allergen traces constitutes a risk to food allergic consumers it is necessary to perform a risk assessment. There are several ways of performing risk assessment for the unintended presence of allergens in food but the underlying need for data is similar (Madsen et al., 2009). Three types of data are needed: The concentration level of the allergen in the food in question, data on the dose that may elicit an allergic reaction and the estimated consumption of the contaminated food. The dose that may elicit a reaction comes from clinical studies where food allergic subjects are challenged with the allergenic food (Taylor et al., 2014). The food producer or a control authority typically provides the level of concentration of the food product.

At the moment in food allergen risk assessment National Food Consumption Surveys are used to estimate consumed quantities for a wide range of food products. However, those surveys are primarily designed to answer nutritional questions, are not performed in a specific food allergic group and the level of detail may not be suitable for the need of food allergen risk assessment. In a peanut allergic population, first investigations had been in the MIRABEL project (Crépet et al., 2015). However, this has limited value for the allergen risk assessment as the food record was performed for limited food groups and not per eating occasion. Moreover, there has been little investigation on how food consumption data can be best used in food allergen risk assessment.

The aim of the work presented in this paper was to develop a standardized, automated procedure to organise data from national food consumption surveys to be suitable to use in food allergen risk assessment. Building on the previous work done by TNO, food groups were proposed using a two steps procedure and using national Food Consumption Surveys from three countries were made available within the iFAAM (integrated approaches to food allergen and allergy management) project: Denmark, France and Netherlands. The procedure was tested on consumption data from three different countries, to investigate if the results were similar and stable across countries.

2. Materials and Methods

In this study we build on the TNO food grouping approach to organize the large amount of food consumption data. Thus, the first step was to organize the food consumption data from the three countries into groups comparable to the original groups from TNO, based on Dutch 2003 consumption data. An automatized two-step procedure was developed building on the unpublished work from TNO describing an approach for food grouping for allergen risk assessment (Figure 1). First, during the clustering method step, the initial groups were automatically split using a common statistics criterion. The aim of the food grouping is to have logical food groups with homogeneous consumption patterns and similar risk outcome. Secondly, the consequences of splitting the initial food groups from the National Food Consumption Survey from each country were investigated. Based on a fixed criterion, we checked if the subdivisions suggested by the clustering method were needed or the difference between risks was too small to make a split meaningful.

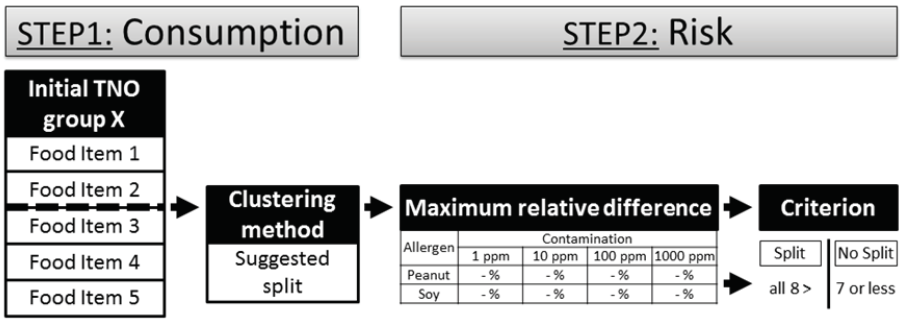


Figure 1: Procedure steps – summary

2.1. National Food Consumption Surveys (Netherlands, France and Denmark)

Detailed food consumption data from the most recent National Food Consumption Survey in Netherlands (van Rossum et al., 2011), France (Dubuisson et al., 2010) and Denmark (Agnes N. Pedersen et al., 2008) were used. Food consumption in the Netherlands was recorded with two non-consecutive 24 hour recalls. In France and Denmark, a pre-coded seven days record was used. We used data from participants aged 18 to 75 years from all three surveys. The food consumption was recorded for each eating occasion, defined as breakfast, lunch, dinner and other distinct eating occasions during the day. The maximum consumption on a

single eating occasion for each consumer and food item was used. As food allergy is an acute reaction, where the allergen if not causing reaction, will not be accumulated, the relevant data are consumption at the meal level (Madsen et al., 2014). All three countries use a common coding system for food items: the Food Ex 2 coding developed by EFSA (European Food Safety Authority, 2015).

2.2. TNO food grouping approach and initial food grouping

TNO has performed allergen risk assessments for several years(Kruizinga et al., 2008; Spanjersberg et al., 2007). Data from the Dutch National Food Consumption Survey 2003 (Ocke et al., 2005) was organised into foods groups. The grouping only includes processed foods based on the assumption that contamination with allergens is not present in raw foods. Furthermore composite dishes were split into their original foods. Based on a standard food grouping used for the Dutch food composition table, consumption data, portion size and if needed, expert judgement, food products were combined into meaningful groups e.g. cookies, ice cream, bread, etc... It resulted in 42 food groups with similar food items and similar intakes, which were validated using allergen risk assessment. These 42 food groups were applied to the food consumption data from the three countries coded at the FoodEx 2 level resulting in the initial food grouping used as a basis for this study.

2.3. Clustering method

The first analysis was to test the differences between food item consumption distributions within each of the initial groups for each country. This was done on the consumption quantities of the three countries separately. The results of these analyses give an indication on how the initial groups can be subdivided to new food groups with similar consumption patterns.

The two non-parametric tests, the Kolmogorov-Smirnov test and the Cramér-van Mises test, commonly used to check if two distributions are drawn from the same distribution, were applied.

Comparing the probability distributions of two food item consumed quantities, the Kolmogorov-Smirnov statistic can be written as follow (Conover, 1971):

$$D_{N,\ M} = \sup_{x} |F_N(x) - G_M(x)|$$

where N and M are the numbers of observations for the empirical distribution functions F_N and G_M of the first and second food items. And sup is the supremum function (maximum).

The Cramér-van Mises criterion is defined as (Anderson, 1962):

$$T = \int_{-\infty}^{\infty} [F_N(x) - G_M(x)]^2 dH_{N+M}(x)$$

where Itm1 and Itm 2 are the numbers of observations for the empirical distribution functions, F_N and G_M of the 1st and 2nd food items. And, $H_{N+M}(x)$ is the empirical distribution function of the 2 food items together.

The KS and Cramér test statistics measure the distance between the food consumption distributions in different ways. The KS test statistics evaluates the difference between distributions based on one point only (the maximum difference, represented by the arrow in Figure 1) whereas the Cramér distance calculates the difference using the area between the two distributions' curves (grey part in Figure 1).

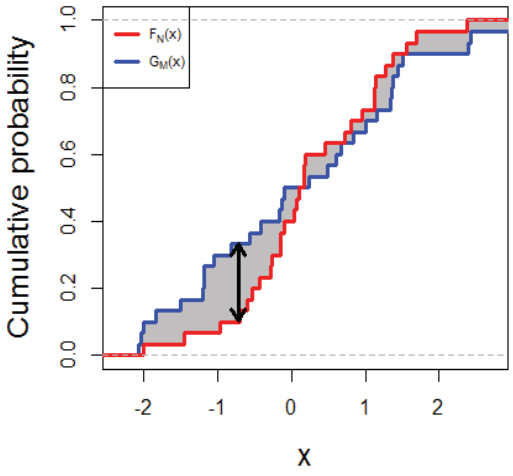


Figure 2: Illustrative example difference between KS and Cramér statistics (the arrow represents the KS statistic and the grey area the Cramér statistic)

Therefore, we calculated the KS and Cramer test statistics for each pair of food items. Two food items with similar distribution will have a small test statistics. And oppositely, the difference in distribution between two

food items will increase with the test statistics value. Thus, the test statistics quantifies the distance between the empirical distribution function of two food items and measures the similarities between food items within each initial group. Using the KS and Cramér test statistics two distances matrices are calculated, they quantify the difference between each food items within each initial group.

Then, the distance matrices are used as an input to the clustering method. It resulted in different suggestion for subdivisions; from not splitting an initial food group to having one food item in each subdivision. The complete-linkage clustering method is used, so the number of food items in each suggested subdivisions is balanced (Sørenson, 1948). Finally, the optimal split is selected based on the average silhouette index (Rousseeuw, 1987). For each subdivision, this index assesses which food items lies together within each subdivision and which food items could be in another subdivision. So, the validity and the quality of the different suggested subdivisions is evaluated and only the subdivisions for which the food items have the most similar consumption patterns are used for the next step of the procedure.

2.4. Risk simulation and clustering validation

Once the optimal split, as described above, was selected the next step was to investigate if the proposed subgrouping is appropriate in terms of estimated risk of an unexpected allergic reaction. This was done by comparing the risk in the subgroups with the risk in the group before subdivision. To calculate the risk a probabilistic risk assessment method was used: the method uses challenge data with allergenic foods, contamination data and consumption data. From these three inputs, the percentage of the population likely to have an allergic reaction is estimated, i.e. the chance for an allergic consumer to eat a high enough amount of allergen to trigger a reaction (Rimbaud et al., 2010; Spanjersberg et al., 2007).

2.4.1. Oral food challenge

Food allergen risk assessment use data where food allergic subjects are challenged in a double blind placebo controlled food challenge (DBPCFC) with the offending food, to determine the dose that elicits an objective allergic reaction. If there is enough data available it is possible to create distributions (Taylor et al., 2014). In

this study, a selection described in Taylor et al. (2014) was used, this selection was based on already published data. The threshold distributions are based on discrete NOAEL (No Observable Adverse Effect Level) and LOAEL (Lowest Observable Adverse Effect Level) values from 158 DBPCFCs for peanut (Anagnostou et al., 2009; Atkins et al., 1985; Blumchen et al., 2010; Clark and Ewan, 2008; Hourihane et al., 1997; Leung et al., 2003; Lewis et al., 2005; NELSON et al., 1997; Nicolaou et al., 2010; Oppenheimer et al., 1992; Patriarca et al., 2006; Wainstein et al., 2010) and 43 for soy (Ballmer-Weber et al., 2007; Fiocchi et al., 2003; Magnolfi et al., 1996; Zeiger et al., 1999). These two allergens were selected because they represent a large difference in challenge data. Data were fitted using the Weibull distribution, a distribution conventionally used to fit allergy thresholds data (Crevel et al., 2007). The Survival package (version 2.38.3) from the R software (version 3.3.0) was used to fit the data (R Core Team, 2015). The fitted distribution was used to calculate the risk of an allergic reaction for the selected food item's distribution. These two threshold distributions are used as a model computational framework for all the food groups and items.

2.4.2. Allergen concentration in food

A range of theoretical concentration data (1, 10, 100 and 1000 ppm protein) were selected to assess the change in risk with increasing contamination. This allowed us to evaluate the behaviour of different quantities of food consumed according to different concentration levels of peanut protein and soy protein.

2.4.3. Risk simulation

As the aim of the analysis was to evaluate the impact of grouping the food items in different food groups according to consumption, the probabilistic risk assessment was performed in a fixed framework of contamination and challenge data. It is assumed that 100% of the products are contaminated and 100% of the consumers were allergic. The simulations were iterated K=1000 times for N=10,000 re-sampling among consumers for each country separately. Thus, for each iteration K, first N consumers were sampled with replacement according to consumers' sampling weights in the surveys. Then, one eating occasion of a food

item among the ones consumed in a food group was randomly sampled. This distribution was used in the risk calculation to represent the consumption in a food groups. Regarding the threshold value, for each iteration K, and each consumer N, a threshold was sampled from the Weibull model presented in the previous section. Finally, the consumption value was multiplied with each unintended allergen level from the range 1, 10, 100 and 1000 ppm.

Thus for each iteration K and each consumer N, we compared the amount of allergens consumed (level of contamination *x* amount consumed) to the simulated thresholds. So for each iteration K, we calculated the percent of allergic reaction among the N consumers. Therefore, for each group and each subgroup, eight different risks were calculated corresponding to the four levels of contamination and the two allergens.

2.4.4 Maximum absolute difference and decision criteria

To assess the impact of difference in consumption on the outcome of the probabilistic risk assessment, the maximum relative difference between the risk of each proposed subgroup and the risk of each initial food group was calculated:

$$max\ relative\ difference\ Group = \max_{Subgroup1,\ subgroup2} \left(\frac{\left| Risk_{Subgroup} - Risk_{\ Group} \right|}{Risk_{Group}} \right)$$

In order to decide whether a group needed to be divided, a formal decision criterion was useful. It helped to assess if the difference between the overall group risk and the subgroup risks is high. Based on similar investigation, no criterion was defined before, thus we chose to develop an objective decision criterion based on the actual data. Thus, based on investigation not shown in this paper, the median of the maximum relative difference on the risk simulated across all the contamination levels, allergen and groups was seen as the best choice to decide if the original groups need to be split or not. Therefore, the criteria were calculated for each country. Groups with a low number of food items, with no subdivisions suggested, was removed from the median calculation. For each group, the criterion was used for all concentration levels and the two allergens.

Therefore, if the risk of all the four concentration levels and the two allergens were above the threshold, it was decided to split the group. Thus, the criterion can be interpreted as such: when the risk connected to consumption of the different subgroups of an initial group is calculated, we can then calculate the maximum relative difference in risk between the overall group's risk and the subgroups' risks. We can then compare them to the criterion to evaluate whether the risks difference between subgroups is so high that we cannot use the whole group consumption distribution to calculate the group risk.

3. Results

3.1. Initial grouping of National Food Consumption Surveys

Table 1 displays the numbers of consumers, the number of food items in the national food consumption surveys and the number of initial grouping. The difference in number of food items reflects the difference in the level of detail when coding foods at the national level.

Country	Number of consumers	Number of food items	Number of groups based on the initial grouping method
Netherlands	3819	496	42
France	2624	367	40
Denmark	2029	168	39

Table 1: Number of consumers and food items in each National Food Consumption Survey and number of groups based on the initial grouping method

3.2. Criterion for each country and example of using the criterion for groups design

The median of the maximum relative difference is 15.1% (Denmark), 13.6% (France) and 16.3% (The Netherlands). This figure for each country was used as the country specific criterion.

Allergen	Contamination			
	1 ppm	10 ppm	100 ppm	1000 ppm
Peanut	46.4%	44.6%	40.3%	31.6%
Soy	48.4%	48.8%	47.1%	44.2%

Table 2 Maximum relative difference in risk between breakfast product group and subgroups in Denmark

The breakfast products' group in Denmark is used to illustrate the grouping design and the application of the method (Table 2). The maximum difference for all the contamination levels and the 2 allergens are higher than the threshold (15.1%). The differences between subgroups are high and this group needs to be split, so single food items need to be allocated to subgroups with more homogeneous consumption patterns.

3.3. Groups outline based on clustering and risk analysis

The first split of the initial groups was based on the clustering method described in the section 2.3. Based on these results each initial group should be divided in one, two or more subgroups. To test the relevance of this grouping to risk assessment the subgroups were evaluated using the risk calculation. Table 3 shows the need for subdivisions based on clustering and the reduced need for subdivision if the analysis is based on risk using the decision criterion. The procedure was performed on the three countries separately.

Number of subdivisions suggested by the clustering	No. of groups needing subdivision based on clustering			No. of groups needing subdivision based on risk assessment analysis		
	Denmark	Netherlands	France	Denmark	Netherlands	France
0	15	7	9	30	25	30
1	23	29	33	9	14	12
2	1	4	0	0	0	0

Table 3: number of subdivisions based on clustering and after the risk assessment analysis for Denmark, Netherlands and France

It was not possible to integrate a criterion to assess the impact of the difference in consumption on the risk during the clustering step. It can be seen that using the risk to evaluate the need of subdividing the initial groups reduces the number of groups that need to be subdivided for the three countries. The procedure minimizes the number of groups with homogeneous consumption patterns.

3.4. Examples with Danish data

Groups of breakfast products, meat, cheese and milk products from the Danish data were chosen to illustrate the procedure. For each group, the risk of an allergic reaction was plotted to illustrate the difference in risk between the subgroups across the concentration levels and for the two allergens. In addition, in order to illustrate the variability within each group and subgroups, the coefficient of variation was calculated (CV= the standard deviation/mean). It measures the dispersion of the consumption around the mean. Those numbers

are reported in a table for each group. Finally, the maximum relative difference for the contamination levels and allergens was also described in a table when it supports the explanation of the choice of splitting or not an initial group.

3.4.1. Splitting the initial group: breakfast products and meat products

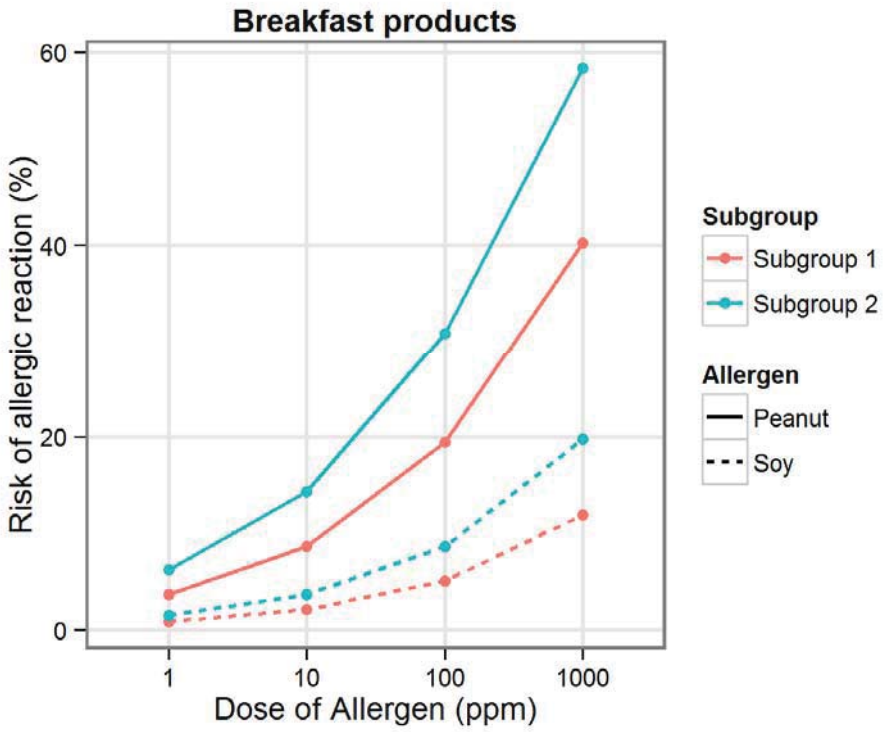


Figure 3: Risk of allergic reaction for two subgroups in the breakfast products group – breakfast product eaten unprocessed (1) and porridge (2)

Breakfast products were divided into two subgroups: breakfast products eaten unprocessed (subgroup 1) and porridges (subgroup 2). There is a large difference in risk between the two subgroups (see Table 2 and Figure 3). The maximum relative differences between the risk of the breakfast products group and its two subgroups for each contamination level and each allergen are all above 15.1%, hence it shows that the difference in consumption between the two subgroups is large, according to the criteria used. Subgroup 2 has a higher consumption (hence risk) than the subgroup 1, so if the combined consumption will be used, it will results in an underestimation of the risk for the subgroup 2 and an overestimation for subgroup 1. Moreover, in Table

4, the coefficient of variation shows that subdividing the breakfast product group also decreases the variability within the consumption subgroups. Thus, the breakfast group is split into two subgroups, so groups with more homogeneous consumption patterns can be used as an input to the risk assessment.

Group of interest	Mean intake	Standard deviation	Coefficient of variation
Overall group	75.0 g	100.0	133.4%
Subgroup 1	44.4 g	28.1	63.3%
Subgroup 2	168.0 g	163.2	97.2%

Table 4: Mean intake in the breakfast products group and subgroups

Figure 4: Risk of allergic reaction for 2 subdivisions in the meat products group – mean consumption 30g (1) and 115g (2)

Group of interest	Mean intake	Standard deviation	Coefficient of variation
Overall group	57.9 g	59.1	102.1%
Subgroup 1	38.3 g	38.8	101.1%
Subgroup 2	111.9 g	70.7	63.2%

Table 5: Mean intake in the meat products group and subgroups

The meat product group was also divided into two. Again, a large difference in risk between the two subgroups can be seen on the Figure 4. In a similar way, dividing this group will help the user to use food distribution with a lower variability for the risk assessment (Table 5).

3.4.2. Keeping the initial group: cheese and milk and milk products

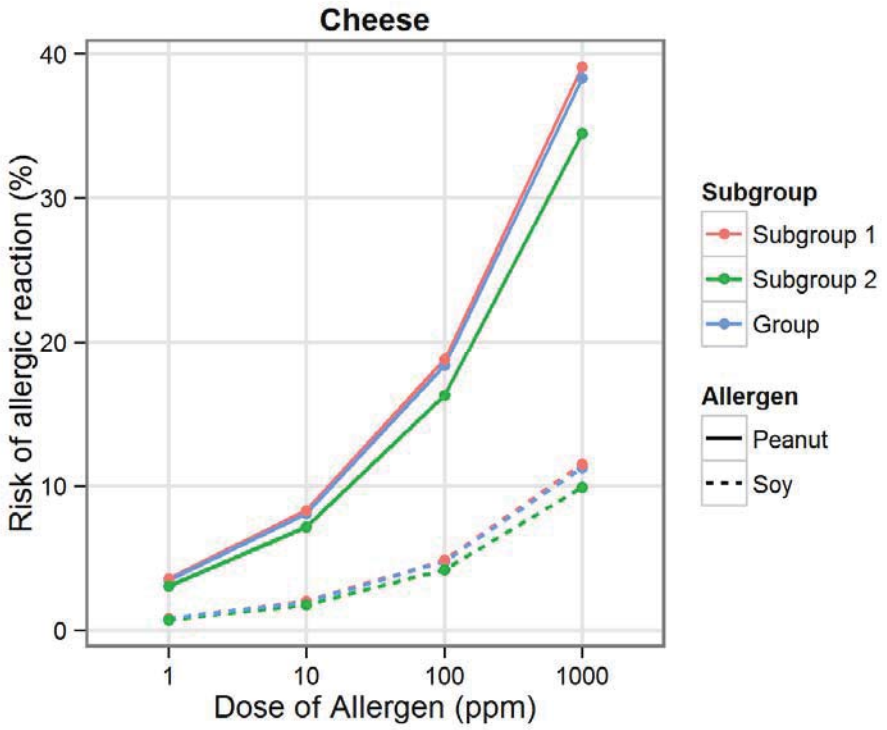


Figure 5: Risk of allergic reaction for the cheese group and for the 2 subdivisions

Allergen	Contamination			
	1 ppm	10 ppm	100 ppm	1000 ppm
Peanut	12.1%	11.9%	11.3%	10.1%
Soy	12.6%	12.2%	12.3%	12.0%

Table 6 Maximum relative difference in risk between cheese group and subgroups in Denmark

Group of interest	Mean intake	Standard deviation	Coefficient of variation
Overall group	37.8 g	26.6	70.4%
Subgroup 1	40.6 g	25.0	61.6%

Subgroup 2	28.4 g	29.8	105.2%

Table 7: Means intake in the cheese group and subgroups

Figure 5 and Table 7 shows that the difference in risk between the two cheese subgroups is very small (all the maximum relative differences are under 15.6%). For this reason the cheese group was not subdivided during the procedure. It is confirmed by the mean consumption for the cheese groups and the two subdivisions: the consumption's means for the food group and its subdivisions are similar. This is confirmed by calculating the coefficient of variation. For the second subgroup, the coefficient of variation is also smaller when the cheese group's consumption distribution is used (Table 7). That's why, there is no reason in splitting the cheese group and then increasing the complexity of the food groups.

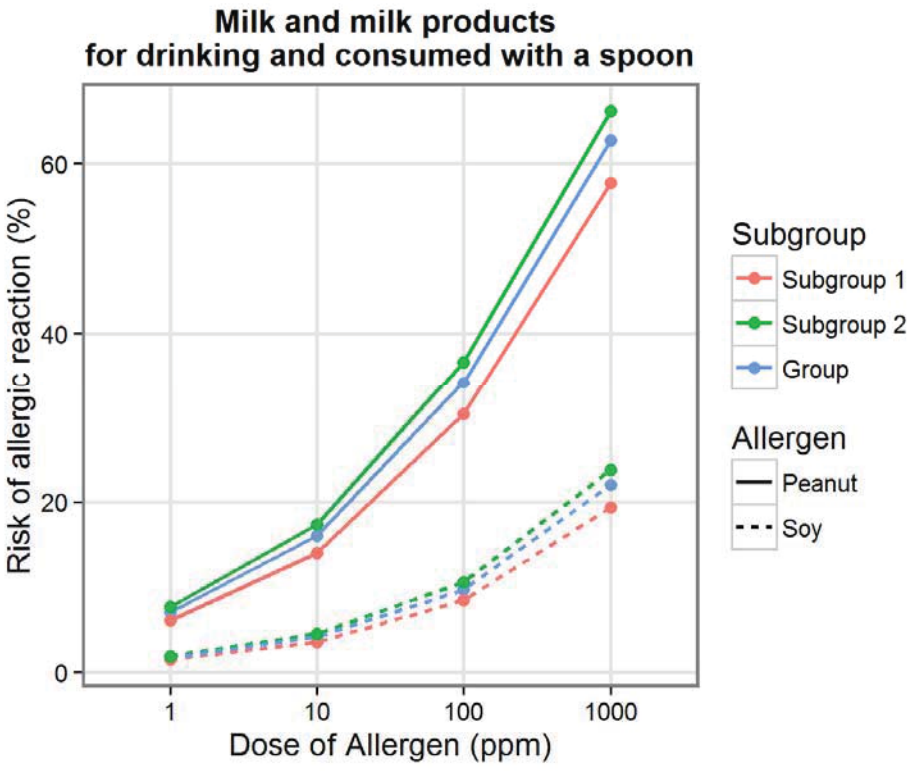


Figure 6: Risk of allergic reaction for the group milk and milk products for its 2 subdivisions

Allergen	Contamination			
	1 ppm	10 ppm	100 ppm	1000 ppm
Peanut	13.0%	12.6%	11.0%	8.1%
Soy	14.1%	13.7%	13.0%	12.1%

Table 8 Maximum relative difference in risk between milk and milk products group and subgroups in Denmark

Group of interest	Mean intake	Standard deviation	Coefficient of variation
Overall group	260.3 g	180.2	69.2%
Subgroup 1	174.0 g	114.6	65.8%
Subgroup 2	324.6 g	192.9	59.4%

Table 9: Mean intake in the milk products group and subgroups

Figure 6 and Table 8 show the group milk products risks and maximum relative differences, the difference in risk between the two subgroups in the milk products group is small for all concentration levels and the two allergens according to the criterion. In Table 9, it can be seen that there is a small difference in consumption in the two subgroups and this difference does not lead to a large difference in risk (according to the criterion). So, the milk and milk products example illustrates the tradeoff between first splitting the group because of the difference in consumption between subgroups within the initial food group based on clustering and the fact that the difference is not high enough according the criterion to split the initial food group. Moreover, the coefficient of variation for the milk product group and its two subgroups are all quite similar, so the homogeneity of consumption patterns will not be improved by splitting this group. Therefore, using the consumption of the whole milk products group in the risk assessment will be adequate to describe it consumption.

4. Discussion/Conclusion

In this paper we present a method to organise food consumption data in groups for use in food allergen risk assessment. The food consumption data from Danish, Dutch and French surveys were initially grouped building on the food grouping approach developed by TNO using the Dutch 2003 survey. The initial groups were validated or split based on the results from the three countries. This resulted in defining 48 food groups in Denmark, 53 in Netherlands and 54 in France. With this study, we provide a method that can be used for any other consumption database from any other country.

In food allergen risk assessment the relevant scenario for consumption is the meal. The data used may include multiple consumptions of the same food from different days. As the aim was to protect the allergic consumers from an adverse reaction, we chose a conservative approach and used the maximum consumption of a food item per person on a single eating occasion. Selecting the maximum consumption assumed a slight overestimation of the risk of allergic reaction, as the average consumption is lower than the one used to estimate the risk of allergic reaction. Rimbaud et al. (2010) also tested random selection within the days of the French National Food Consumption Survey and the difference in risk was not found significant. So, we can expect that using the maximum consumption has a relative low impact on the final group design for each country.

The initial food grouping approach developed by TNO could not be directly used for consumption data from other countries because they were based on the previous Dutch National Food Consumption Surveys and Dutch food codes, which limited the applicability for applying the method to other national food databases, and comparing the food consumption patterns of countries. As the amount of data from national consumption surveys are very large we wanted to develop a method that was more automated based on fixed criteria. However, clustering was based on consumption level and heterogeneity within each initial group, but not comparing the size of the effects across food groups. So, depending on the homogeneity of the initial group, the difference in consumption between subdivisions across initial food groups can be small or high. The second step was designed to investigate the relevance in relation to risk of splitting the groups. This was done comparing the risk between the original and the new groups formed because of subdivision. During the risk analysis, the criteria choice was made so that no subjective information had influence, so there was no bias in deciding if a difference in risk is high. As there were no similar work about evaluating the magnitude of those differences, other measures than the maximum relative difference between the group split and the subdivisions risks were investigated, like the maximum absolute difference. But, some tests (not presented here) showed that, maximum relative difference was the most stable criterion across the condition

to investigate the consequences of subdividing food groups on the risk of allergic reaction. The choice of the median across the allergens and the contamination levels was found to best fit the purpose of having groups with homogeneous consumption patterns with a significant difference in risk. Furthermore, it was decided to use national criteria, as the criteria are so alike for the three countries that there was no reason to try to develop a common criterion.

The Weibull distribution was used to fit the dose response curve from the challenge data. Crevel et al. (2007) propose three distributions for fitting threshold data: Weibull, log-logistic and log-normal. It is assumed that selecting another threshold distribution would influence the outcome of the risk assessment with a comparable factor, so the results of the food grouping will be similar. Due to high amount of data and the number of comparison made in the procedure, it was not realistic to investigate the food group design using the three distributions.

By joining the food items because of the reasons mentioned previously, we choose not to perform the risk assessment at the food items level. It is still possible to perform the risk assessment at the food item level, but it was not the aim of the procedure. At the same time, the procedure also increases the number of consumption data points used to assess the risk of allergic reaction, because several food items in a food group can be used to estimate the risk.

The fixed framework of concentration levels and allergen threshold distribution controlled the way that the differences in consumption propagated to the risk. So, the difference in risks reflected the difference in consumption, as the concentration levels and thresholds have similar impact on the risk calculation, only the consumption distribution varies from the different groups and subgroups. The whole food consumption distribution was used in the risk assessment. So, the variability in the consumption distribution was taken into account, and the spread of the consumption distribution is expected not to influence the food grouping. Moreover, uncertainty on the parameters of the concentration and threshold distribution can be integrated

during the risk calculation (Kruizinga et al., 2008; Rimbaud et al., 2010; Spanjersberg et al., 2007). Thus, the risk could also be estimated with a confidence interval. It was decided not to investigate the impact of those uncertainties, as they were not the focus of this study and it was preferred to evaluate the impact of the difference in consumption on the risk.

The procedure described in the paper is designed in a way that it can be easily applied to any National Food consumption Survey in Europe. For each country the number of suggested subdivisions decreases by around 60% after the risk analysis. So, we assume that food groups based on National Food Consumption Surveys from other countries will have similar number of groups if using the suggested procedure.

Because of similarity in the data among the three countries we have only used data from Denmark in the detailed examples and present the final grouping in Appendix 1 and 2 for Denmark only.

In this paper we present a method that can be used to group food consumption data from one country using the national food consumption survey. In a world were food is exported to many countries it is relevant to be able to do risk assessment on data that cover multiple countries, ideally e.g. the whole of EU. We have therefore investigated how food consumption data from multiple countries can be combined using data from the three countries. These results will be presented in a later paper.

Acknowledgements

This work was part of the iFAAM project (Integrated Approaches to Food Allergen and Allergy Risk Management, Grant Agreement No. 322147), the national food consumption data for France and Denmark were kindly provided by the work package partners: ANSES and DTU. The national food consumption data of the Netherlands was kindly provided by the National Institute for Public Health and the Environment in the Netherlands. The challenge data for peanut and soy was kindly provided by TNO and FARRP (University of

Nebraska – Food Allergy Research and Resource Program). We also thank TNO for providing the information on the food grouping performed on the Dutch data

Appendix 1: Food intake summary statistics per food group in Denmark (in g)

Final food group	Mean	SD	P50	P55	P60	P65	P70	P75	P80	P85	P90	P95	P975	P99	P100
Peanuts, nuts and dried fruits	34.1	29.4	30.0	30.0	30.0	30.0	30.0	40.0	40.0	60.0	60.0	90.0	120.0	150.0	330.0
Potato and other starch based chips (including salty sticks)	42.1	38.6	30.0	30.0	30.0	45.0	45.0	50.0	60.0	75.0	90.0	150.0	150.0	181.2	255.0
Fried/warm snacks	162.8	103.1	150.0	150.0	180.0	180.0	180.0	180.0	225.0	270.0	270.0	360.0	448.1	450.0	720.0
Meal replacements and meat imitates	40.1	18.7	32.0	32.0	32.0	38.4	51.2	64.0	64.0	64.0	64.0	64.0	64.0	64.0	64.0
Pancakes and waffles	151.5	104.6	100.0	150.0	150.0	150.0	150.0	200.0	200.0	250.0	300.0	335.0	400.0	500.0	800.0
Soups	395.9	169.8	375.0	500.0	500.0	500.0	500.0	500.0	500.0	500.0	500.0	750.0	750.0	750.0	1000.0
Small sweets	41.1	39.4	25.0	25.0	25.0	50.0	60.0	60.0	60.0	60.0	100.0	100.0	120.0	200.0	420.0
Sugar	21.4	12.9	16.0	16.0	22.0	23.0	23.0	24.0	32.0	32.0	33.0	46.0	56.8	68.0	120.0
Peanut butter	30.0	27.0	23.0	23.0	24.0	30.0	32.0	32.0	39.0	45.0	60.0	75.0	97.5	111.0	120.0
Chocolate and chocolate products	43.5	34.2	25.0	32.6	50.0	60.0	60.0	60.0	60.0	60.0	100.0	100.0	120.0	180.0	400.0
Sweet confectionary (jam, marmalade)	19.8	13.1	16.0	16.0	16.0	24.0	24.0	24.0	24.0	32.0	32.0	40.0	56.0	80.0	80.0
Cereal bars	27.2	24.2	25.0	25.0	25.0	25.0	25.0	25.0	25.0	32.4	45.0	50.0	50.4	92.6	250.0
Chewing gum	10.6	7.8	10.0	10.0	10.0	10.0	10.0	10.0	15.0	15.0	20.0	25.0	30.0	37.9	60.0
Mashed potato powder	151.3	87.9	139.0	139.0	139.0	139.0	139.0	212.0	222.0	222.0	222.0	340.0	436.2	444.0	680.0
Potato product (excl. powder)	221.3	120.9	225.0	225.0	225.0	250.0	300.0	300.0	300.0	375.0	375.0	434.0	464.0	530.0	1025.0
Vegetable oils and animal fat	21.0	16.7	16.0	16.0	20.0	24.0	24.0	30.0	32.0	32.0	45.0	49.0	64.0	80.0	150.0
Vegetable oils and animal fat	16.0	13.6	12.0	15.0	16.0	16.0	16.0	20.0	24.0	28.0	32.0	40.0	49.2	64.0	176.0
Sauces used as condiments and dessert sauces	26.1	22.4	15.0	21.0	30.0	30.0	30.0	30.0	30.0	45.0	45.0	60.0	75.0	116.1	450.0
Sauces , savory, chutneys and pickles	68.6	49.3	60.0	60.0	60.0	75.0	75.0	90.0	90.0	105.0	120.0	150.0	180.0	300.0	600.0
Cheese	37.8	26.6	32.0	32.0	32.0	40.0	46.0	46.0	48.0	64.0	69.0	90.6	96.0	134.4	300.0
Fish products – mean consumption 30 g[a]	30.8	25.4	23.0	23.0	23.0	30.0	32.0	40.0	40.0	46.0	60.0	80.0	102.5	127.5	160.0
Fish products - mean consumption 115 g[b]	122.9	110.6	100.0	101.0	101.0	110.0	126.0	150.0	162.0	200.1	202.0	303.0	324.0	458.5	1134.0
Meat products – mean consumption 65 g[c]	38.3	38.8	32.0	32.0	32.0	32.0	40.0	46.0	46.0	64.0	70.0	98.1	140.0	210.0	560.0
Meat products -mean consumption 105 g[d]	111.9	70.7	101.0	101.0	101.0	110.0	110.0	126.0	150.0	162.0	202.0	243.0	301.1	330.0	880.0
Crackers, crisp bread, rusk and toast	24.0	15.2	24.0	24.0	24.0	24.0	24.0	24.0	24.0	36.0	36.0	48.0	60.0	72.0	180.0
Bread, bread rolls and bread doughs	92.6	47.5	80.0	86.0	90.0	98.0	120.0	120.0	120.0	130.0	160.0	180.0	199.5	240.0	510.0
Herbs and spices mixes, bouillon cubes, yeast extract	18.1	32.4	10.0	10.0	10.0	15.0	20.0	20.0	20.0	20.0	20.0	35.5	65.5	200.0	400.0
Alcoholic drinks, alcohol ≤15%	272.8	149.9	280.0	280.0	280.0	280.0	280.0	280.0	420.0	420.0	420.0	560.0	633.5	700.0	1120.0

Alcoholic drinks,alcohol above 15%	61.8	50.6	50.0	60.0	60.0	60.0	60.0	60.0	90.0	90.0	100.0	150.0	180.0	240.0	699.0
Beer	672.8	624.4	330.0	495.0	660.0	660.0	660.0	660.0	990.0	990.0	1320.0	1980.0	2640.0	3300.0	5940.0
Drinks without alcohol (excl. syrup)	479.6	291.8	400.0	400.0	500.0	540.0	600.0	600.0	600.0	800.0	800.0	1000.0	1200.0	1600.0	5000.0
Cookies (biscuits)	37.6	25.5	30.0	36.0	36.0	36.0	47.5	48.0	48.0	60.0	72.0	96.0	120.0	120.0	192.0
Macaroons	14.9	13.1	7.0	14.0	14.0	14.0	14.0	14.0	21.0	22.4	28.0	44.8	56.0	57.1	70.0
Cakes (including pastry)	137.2	69.5	125.0	125.0	125.0	125.0	125.0	125.0	200.0	200.0	250.0	250.0	300.0	375.0	875.0
Breakfast products eaten unprocessed (e.g. müsli, oat and maize flakes)	44.4	28.1	46.0	46.0	46.0	46.0	47.0	64.0	64.0	64.0	92.0	92.0	117.0	117.0	234.0
Breakfast products, porridge	168.0	163.2	150.0	150.0	150.0	150.0	150.0	202.0	202.0	225.0	257.0	300.0	356.0	502.9	2850.0
Maize grain	56.2	45.7	38.0	38.0	38.0	60.0	64.0	64.0	64.0	114.1	115.0	119.6	215.0	215.0	345.0
Pasta, rice, couscous and other grains	159.7	102.0	125.0	139.0	139.0	148.0	208.0	208.0	222.0	222.0	278.0	363.0	416.0	444.0	726.0
Legumes[e]	109.5	70.4	82.0	106.0	106.0	106.0	106.0	130.8	139.0	139.0	215.0	230.4	282.7	363.9	444.0
Fruit and vegetables, processed	133.0	86.1	115.0	115.0	135.0	135.0	159.0	185.0	200.0	215.0	238.0	238.0	370.0	476.0	714.0
Eggs	42.1	28.7	40.0	40.0	46.0	55.0	55.0	55.0	55.0	55.0	80.0	110.0	110.0	110.0	165.0
Egg based dishes (Omelette plain, fried eggs)	68.2	38.0	50.0	50.0	50.0	79.1	100.0	100.0	100.0	100.0	100.0	100.0	150.0	200.0	300.0
Omellette with bacon	149.3	69.0	100.0	110.5	200.0	200.0	200.0	200.0	200.0	200.0	200.0	245.0	300.0	400.0	600.0
Sandwich and pizza	366.8	263.4	290.0	300.0	350.0	400.0	450.0	450.0	600.0	600.0	900.0	900.0	900.0	900.0	2000.0
Composite dishes[f]	204.9	122.1	165.0	165.0	208.0	237.0	237.0	237.0	248.0	330.0	358.0	450.0	474.0	716.0	1080.0
Cream and coffee milk	18.8	25.9	14.0	14.0	14.0	14.0	20.3	21.0	21.0	28.0	28.0	28.0	32.6	110.7	199.5
Ice cream	77.7	45.9	50.0	50.0	75.0	100.0	100.0	100.0	100.0	100.0	100.0	150.0	200.0	200.0	600.0
Milk and milk products for drinking and consumed with a spoon	260.3	180.2	200.0	200.0	200.0	200.0	250.0	400.0	400.0	400.0	400.0	600.0	600.0	800.0	3000.0

[a] such as: prepared fish salad, fish fingers breaded, canned seafood

[b] such as: fish balls

[c] such as: bacon, cooked sausages, pat

[d] such as: meat loaf, meat balls, meat burger

[e] such as: dry seeds

[f] such as: lasagne, goulash, fish gratin

Appendix 2: Final Danish food groups and the overall FoodEx 2 groups included

Final food groups	FoodEx 2 Code	Foods included in final group (FoodEx 2 Name)
Peanut butter	A01BN	Peanut butter
Cheese	A02QE	Cheese
Cheese	A02QF	Fresh uncured cheese
Cheese	A02TQ	Cheese, danbo
Cheese	A031A	Processed cheese and spreads
Cream and coffee milk	A02ML	Cream, plain
Ice cream	A02QA	Ice cream, milk-based
Ice cream	A036J	Water-based ice creams
Milk and milk products for drinking and consumed with a spoon	A02LV	Cow milk
Milk and milk products for drinking and consumed with a spoon	A02LY	Cow milk, whole
Milk and milk products for drinking and consumed with a spoon	A02LZ	Cow milk, semi skimmed (half fat)
Milk and milk products for drinking and consumed with a spoon	A02MA	Cow milk, skimmed (low fat)
Milk and milk products for drinking and consumed with a spoon	A02MP	Flavoured milks
Milk and milk products for drinking and consumed with a spoon	A02NG	Yoghurt, cow milk, plain
Milk and milk products for drinking and consumed with a spoon	A02NH	Yoghurt, cow milk, flavoured
Milk and milk products for drinking and consumed with a spoon	A02NQ	Yoghurt drinks, sweetened and/or flavoured
Milk and milk products for drinking and consumed with a spoon	A02QD	Milkshakes
Milk and milk products for drinking and consumed with a spoon	A03TJ	Soya drink
Milk and milk products for drinking and consumed with a spoon	A03TM	Rice drink
Milk and milk products for drinking and consumed with a spoon	A041E	Rice pudding
Milk and milk products for drinking and consumed with a spoon	A04NS	Other desserts spoonable
Peanuts, nuts and dried fruits	A014C	Tree nuts
Peanuts, nuts and dried fruits	A014K	Coconut
Peanuts, nuts and dried fruits	A015F	Oilseeds
Peanuts, nuts and dried fruits	A015H	Peanut
Peanuts, nuts and dried fruits	A01MA	Dried fruit
Peanuts, nuts and dried fruits	A01MB	Dried prunes
Peanuts, nuts and dried fruits	A01MD	Dried apricots
Peanuts, nuts and dried fruits	A01MF	Dried dates
Peanuts, nuts and dried fruits	A01MG	Dried figs
Peanuts, nuts and dried fruits	A01QF	Mixed dried fruits
Potato and other starch based chips (including salty sticks)	A00DC	Popcorn (maize, popped)
Potato and other starch based chips (including salty sticks)	A011L	Potato crisps (chips)
Fried/warm snacks	A040D	Sausage roll
Fried/warm snacks	A040F	Spring rolls
Meal replacements and meat imitates	A03RS	Food for weight reduction
Pancakes and waffles	A00CL	Pancakes
Soups	A041S	Other vegetable soup
Soups	A041V	Meat and vegetable soup
Soups	A042A	Fruit soup
Small sweets	A034V	Sweet confectionery
Small sweets	A034X	Candies
Small sweets	A034Y	Marzipan
Small sweets	A035H	Foamed sugar products (marshmallows)
Small sweets	A035J	Liquorice candies
Small sweets	A035K	Gum drops
Sugar	A032H	Sucrose (common sugar)
Sugar	A04PA	Sugar and other sweetening ingredients

		(excluding intensive sweeteners)
Chocolate and chocolate products	A034F	Chocolate
Sweet confectionary (jam, marmalade)	A01MM	Jam
Cereal bars	A00EY	Cereal bars
Cereal bars	A00FH	Mixed cereal-based snacks
Chewing gum	A035M	Chewing gum
Mashed potato powder	A011E	Mashed potato powder
Potato product (excl. powder)	A011P	Potato boiled
Potato product (excl. powder)	A011R	Potato baked
Potato product (excl. powder)	A03VD	Potato based dishes
Potato product (excl. powder)	A03VG	Prepared potato salad
Potato product (excl. powder)	A0BYV	French fries from cut potato
Vegetable oils and animal fat	A036P	Olive oils
Vegetable oils and animal fat	A036V	Rape seed oil, edible
Vegetable oils and animal fat	A037V	Pork lard
Vegetable oils and animal fat	A039E	Blended margarine
Vegetable oils and animal fat	A04SD	Blended fat and oils
Vegetable oils and animal fat	A037V	Pork lard
Butter/halvarine/margarine	A039C	Butter
Butter/halvarine/margarine	A039D	Traditional margarine
Sauces , savory, chutneys and pickles	A043V	Savoury sauces
Sauces , savory, chutneys and pickles	A044E	Vegetables-based coocked sauce
Sauces used as condiments and dessert sauces	A044F	Table-top condiments
Sauces used as condiments and dessert sauces	A044R	Soy sauce
Sauces used as condiments and dessert sauces	A044V	Pesto
Sauces used as condiments and dessert sauces	A044X	Mayonnaise
Sauces used as condiments and dessert sauces	A045J	Mixed and other not listed condiments
Sauces used as condiments and dessert sauces	A045K	Salad dressing
Sauces used as condiments and dessert sauces	A045L	Salad dressing, low fat
Sauces used as condiments and dessert sauces	A045N	Tartar sauce
Fish products - mean 30 g	A02KC	Fish fingers, breaded
Fish products - mean 30 g	A03XP	Prepared fish salad
Fish products - mean 30 g	A0BZ5	Canned seafood
Fish products - mean 110 g	A02KD	Fish balls
Meat products - mean 65 g	A022X	Bacon
Meat products - mean 65 g	A023G	Cooked cured meat
Meat products - mean 65 g	A025J	Cooked sausages
Meat products - mean 65 g	A025Z	Head cheese
Meat products - mean 65 g	A026R	Pate, pork liver
Meat products - mean 65 g	A04NG	Sausages and other comminuted meat
Meat products -mean 105 g	A026D	Wiener sausage
Meat products -mean 105 g	A03VV	Meat based dishes
Meat products -mean 105 g	A03XA	Meat loaf
Meat products -mean 105 g	A03XD	Meat loaf with cheese, vegetables or other
Meat products -mean 105 g	A03XF	Meat burger (no sandwich)
Meat products -mean 105 g	A03XG	Meat balls
Crackers, crisp bread, rusk and toast	A006D	Wheat crisp bread
Bread, bread rolls and bread doughs	A004V	Bread and similar products
Bread, bread rolls and bread doughs	A004Y	Wheat bread and rolls, white (refined flour)
Bread, bread rolls and bread doughs	A005E	Wheat bread and rolls, brown or wholemeal
Bread, bread rolls and bread doughs	A005H	Rye bread, wholemeal
Bread, bread rolls and bread doughs	A006S	Pita bread
Bread, bread rolls and bread doughs	A006V	Tortilla
Herbs and spices mixes, bouillon cubes, yeast extract	A01AM	Caper buds

Herbs and spices mixes, bouillon cubes, yeast extract	A042Y	Seasoning mixes
Herbs and spices mixes, bouillon cubes, yeast extract	A043H	Stock cube, beef flavour
Alcoholic drinks, alcohol ≤15%	A03MV	Wine, white
Alcoholic drinks, alcohol ≤15%	A03MX	Wine, red
Alcoholic drinks, alcohol ≤15%	A03ND	Cider
Alcoholic drinks, alcohol above 15%	A03NG	Fortified and liqueur wines
Alcoholic drinks, alcohol above 15%	A03PD	Unsweetened spirits
Alcoholic drinks, alcohol above 15%	A04QG	Other spirits
Beer	A03MD	Beer, strong
Beer	A03ME	Beer, regular
Beer	A03MF	Beer, light
Drinks without alcohol (excl. syrup)	A03KC	Coffee (average strenght) beverage
Drinks without alcohol (excl. syrup)	A03KG	Coffee with milk or cream
Drinks without alcohol (excl. syrup)	A03LB	Tea beverages
Drinks without alcohol (excl. syrup)	A03KH	Coffee drink, cappuccino
Drinks without alcohol (excl. syrup)	A03LG	Herbal and other non-tea infusions
Drinks without alcohol (excl. syrup)	A039L	Fruit juices
Drinks without alcohol (excl. syrup)	A03CH	Vegetable juices, ready to drink
Drinks without alcohol (excl. syrup)	A04PX	Bottled water
Drinks without alcohol (excl. syrup)	A03DZ	Soft drinks
Drinks without alcohol (excl. syrup)	A03FT	Diet soft drinks
Cookies (biscuits)	A009V	Biscuits (sweet and semi-sweet)
Cookies (biscuits)	A009X	Biscuits, sweet, plain
Cookies (biscuits)	A00AE	Biscuit, filled (with inclusions, filling or coating)
Macaroons	A00CN	Macaroons
Cakes (including pastry)	A00AV	Cream cake
Cakes (including pastry)	A00BZ	Fruit pie-tarts
Cakes (including pastry)	A00CC	Puff pastry
Breakfast products eaten unprocessed (e.g. müsli, oat and maize flakes)	A00DH	Oat rolled grains
Breakfast products eaten unprocessed (e.g. müsli, oat and maize flakes)	A00EJ	Muesli and similar
Breakfast products eaten unprocessed (e.g. müsli, oat and maize flakes)	A04QY	Other processed and mixed breakfast cereals
Breakfast products, porridge	A00EN	Porridge
Breakfast products, porridge	A00ES	Rye porridge
Maize grain	A000T	Maize grain
Pasta, rice, couscous and other grains	A003E	Rice grain, parboiled
Pasta, rice, couscous and other grains	A004G	Bulgur
Pasta, rice, couscous and other grains	A007D	Pasta and similar products
Pasta, rice, couscous and other grains	A040Z	Rice based dishes
Legumes	A012R	Pulses (dry)
Fruit and vegetables, processed	A01QG	Fruit salad
Fruit and vegetables, processed	A03YE	Mixed vegetables, boiled
Fruit and vegetables, processed	A03YF	Vegetables, gratinated
Fruit and vegetables, processed	A042B	Salads
Fruit and vegetables, processed	A042F	Greek salad
Eggs	A031S	Hen egg yolk
Eggs	A032B	Boiled eggs
Egg based dishes (Omelette plain, fried eggs)	A032C	Fried eggs
Egg based dishes (Omelette plain, fried eggs)	A03YN	Omelette, plain
Omellette with bacon	A03YP	Omelette with bacon
Sandwich and pizza	A03ZA	Sandwich with cheese topping/filling
Sandwich and pizza	A03ZG	Sandwich with meat and vegetable topping/filling
Sandwich and pizza	A03ZK	Hot dog with bread
Sandwich and pizza	A03ZL	Hamburger with bread
Sandwich and pizza	A040B	Pizza with cheese, meat, mushrooms, and vegetables

Composite dishes	A03VK	Potatoes, meat, and vegetables meal
Composite dishes	A03VS	Beans and vegetables meal
Composite dishes	A03VX	Goulash
Composite dishes	A03VY	Meat stew
Composite dishes	A03XJ	Fish and seafood based dishes
Composite dishes	A03XL	Fish gratin
Composite dishes	A03XM	Seafood-based meals
Composite dishes	A03XS	Fish and vegetables meal
Composite dishes	A03XZ	Vegetable casserole
Composite dishes	A03YA	Veggie pot pie
Composite dishes	A03YM	Quiche
Composite dishes	A040P	Lasagna
Composite dishes	A041J	Rice, meat, and vegetables meal

Bibliography

Agnes N. Pedersen, Fagt, S., Groth, M.V., Christensen, T., Biltoft-Jensen, A., Matthiessen, J., Andersen, N.L., Kørup, K., Hartkopp, H., Ygil, K.H., Hinsch, H.-J., Saxholt, E., Trolle, E., 2008. Danskernes kostvaner 2003-2008, Danskernes kostvaner 2003-2008.

Anagnostou, K., Islam, S., King, Y., Deighton, J., Clark, A.T., Ewan, P.W., 2009. British Society for Allergy and Clinical Immunology Annual Conference 2009 Abstracts, in: British Society for Allergy and Clinical Immunology Annual Conference 2009 Abstracts. Blackwell Publishing Ltd, pp. 1937–1958. doi:10.1111/j.1365-2222.2009.03389.x

Anderson, T.W., 1962. On the Distribution of the Two-Sample Cramer-von Mises Criterion 1148–1159. doi:10.1214/aoms/1177704477

Atkins, F.M., Steinberg, S.S., Metcalfe, D.D., 1985. Evaluation of immediate adverse reactions to foods in adult patients: II. A detailed analysis of reaction patterns during oral food challenge. J. Allergy Clin. Immunol. 75, 356–363. doi:10.1016/0091-6749(85)90072-7

Ballmer-Weber, B.K., Holzhauser, T., Scibilia, J., Mittag, D., Zisa, G., Ortolani, C., Oesterballe, M., Poulsen, L.K., Vieths, S., Bindslev-Jensen, C., 2007. Clinical characteristics of soybean allergy in Europe: A double-blind, placebo-controlled food challenge study. J. Allergy Clin. Immunol. 119, 1489–1496. doi:10.1016/j.jaci.2007.01.049

Blumchen, K., Ulbricht, H., Staden, U., Dobberstein, K., Beschorner, J., de Oliveira, L.C.L., Shreffler, W.G., Sampson, H.A., Niggemann, B., Wahn, U., Beyer, K., 2010. Oral peanut immunotherapy in children with peanut anaphylaxis. J. Allergy Clin. Immunol. 126, 83–91.e1. doi:10.1016/j.jaci.2010.04.030

Bock, S.A., Muoz-Furlong, A., Sampson, H.A., 2001. Fatalities due to anaphylactic reactions to foods. J. Allergy Clin. Immunol. 107, 191–193. doi:10.1067/mai.2001.112031

Clark, A.T., Ewan, P.W., 2008. Good prognosis, clinical features, and circumstances of peanut and tree nut reactions in children treated by a specialist allergy center. J. Allergy Clin. Immunol. 122, 286–289. doi:10.1016/j.jaci.2008.05.015

Conover, W.J., 1971. Practical Nonparametric Statistics. Stat. doi:10.2307/2986830

Crépet, A., Papadopoulos, A., Elegbede, C.F., Ait-Dahmane, S., Loynet, C., Millet, G., Van Der Brempt, X., Bruyère, O., Marette, S., Moneret-Vautrin, D.A., 2015. Mirabel: An integrated project for risk and cost/benefit analysis of peanut allergy. Regul. Toxicol. Pharmacol. 71, 178–183. doi:10.1016/j.yrtph.2014.12.006

Crevel, R.W.R., Briggs, D., Hefle, S.L., Knulst, A.C., Taylor, S.L., 2007. Hazard characterisation in food allergen risk assessment: the application of statistical approaches and the use of clinical data. Food Chem. Toxicol. 45, 691–701. doi:10.1016/j.fct.2006.09.005

Dubuisson, C., Lioret, S., Touvier, M., Dufour, A., Calamassi-Tran, G., Volatier, J.-L., Lafay, L., 2010. Trends in food and nutritional intakes of French adults from 1999 to 2007: results from the INCA surveys. Br. J. Nutr. 103, 1035–48. doi:10.1017/S0007114509992625

DunnGalvin, A., Chan, C.H., Crevel, R., Grimshaw, K., Poms, R., Schnadt, S., Taylor, S.L., Turner, P., Allen, K.J., Austin, M., Baka, A., Baumert, J.L., Baumgartner, S., Beyer, K., Bucchini, L., Fernández-Rivas, M., Grinter, K., Houben, G.F., Hourihane, J., Kenna, F., Kruizinga, A.G., Lack, G., Madsen, C.B., Clare Mills, E.N., Papadopoulos, N.G., Alldrick, A., Regent, L., Sherlock, R., Wal, J.M., Roberts, G., 2015. Precautionary allergen labelling: Perspectives from key stakeholder groups. Allergy Eur. J. Allergy Clin. Immunol. 70, 1039–1051. doi:10.1111/all.12614

European Food Safety Authority, 2015. The food classification and description system FoodEx 2 (revision 2). EFSA Support. Publ. 8, n/a–n/a. doi:10.2903/sp.efsa.2011.EN-215

Fernández-Rivas, M., Asero, R., 2014. Risk Management for Food Allergy, Risk Management for Food Allergy. Elsevier. doi:10.1016/B978-0-12-381988-8.00002-6

Fiocchi, A., Travaini, M., D'Auria, E., Banderali, G., Bernardo, L., Riva, E., 2003. Tolerance to a rice hydrolysate formula in children allergic to cow's milk and soy. Clin. Exp. Allergy 33, 1576–1580. doi:10.1046/j.1365-2222.2003.01781.x

Hourihane, J.O., Kilburn, S.A., Nordlee, J.A., Hefle, S.L., Taylor, S.L., Warner, J.O., 1997. An evaluation of the sensitivity of subjects with peanut allergy to very low doses of peanut protein: A randomized, double-blind, placebo-controlled food challenge study. J. Allergy Clin. Immunol. 100, 596–600. doi:10.1016/S0091-6749(97)70161-1

Kruizinga, A.G., Briggs, D., Crevel, R.W.R., Knulst, A.C., Bosch, L.M.C. van den, Houben, G.F., 2008. Probabilistic risk assessment model for allergens in food: sensitivity analysis of the minimum eliciting dose and food consumption. Food Chem. Toxicol. 46, 1437–1443. doi:10.1016/j.fct.2007.09.109

Leung, D.Y.M., Sampson, H.A., Yunginger, J.W., Burks, A.W., Schneider, L.C., Wortel, C.H., Davis, F.M., Hyun, J.D., Shanahan, W.R., 2003. Effect of Anti-IgE Therapy in Patients with Peanut Allergy. N. Engl. J. Med. 348, 986–993. doi:10.1056/NEJMoa022613

Lewis, S.A., Grimshaw, K.E.C., Warner, J.O., Hourihane, J.O., 2005. The promiscuity of immunoglobulin E binding to peanut allergens, as determined by Western blotting, correlates with the severity of clinical symptoms. Clin. Exp. Allergy 35, 767–773. doi:10.1111/j.1365-2222.2005.02252.x

Madsen, C.B., Hattersley, S., Buck, J., Gendel, S.M., Houben, G.F., Hourihane, J.O., Mackie, A., Mills, E.N.C., Nørhede, P., Taylor, S.L., Crevel, R.W.R., 2009. Approaches to risk assessment in food allergy: report from a workshop developing a framework for assessing the risk from allergenic foods". Food Chem. Toxicol. 47, 480–9. doi:10.1016/j.fct.2008.12.001

Madsen, C.B., Houben, G., Hattersley, S., Crevel, R.W.R., Remington, B.C., Baumert, J.L., 2014. Risk Management for Food Allergy, Risk Management for Food Allergy. Elsevier. doi:10.1016/B978-0-12-381988-8.00006-3

Magnolfi, C.F., Zani, G., Lacava, L., Patria, M.F., Bardare, M., 1996. Soy Allergy in Atopic Children. Ann. Allergy, Asthma Immunol. 77, 197–201. doi:10.1016/S1081-1206(10)63255-3

NELSON, H., LAHR, J., RULE, R., BOCK, A., LEUNG, D., 1997. Treatment of anaphylactic sensitivity to peanuts by immunotherapy with injections of aqueous peanut extract1. J. Allergy Clin. Immunol. 99, 744–751. doi:10.1016/S0091-6749(97)80006-1

Nicolaou, N., Poorafshar, M., Murray, C., Simpson, A., Winell, H., Kerry, G., Härlin, A., Woodcock, A., Ahlstedt,

S., Custovic, A., 2010. Allergy or tolerance in children sensitized to peanut: Prevalence and differentiation using component-resolved diagnostics. J. Allergy Clin. Immunol. 125, 191–197.e13. doi:10.1016/j.jaci.2009.10.008

Nwaru, B.I., Hickstein, L., Panesar, S.S., Roberts, G., Muraro, A., Sheikh, A., 2014. Prevalence of common food allergies in Europe: A systematic review and meta-analysis. Allergy Eur. J. Allergy Clin. Immunol. 69, 992–1007. doi:10.1111/all.12423

Ocke, M., Hulshof, K., van Rossum, C., 2005. The Dutch National Food Consumption Survey - 2003, Methodological issues. Arch. Public Heal. 63, 227–241.

Oppenheimer, J.J., Nelson, H.S., Bock, S.A., Christensen, F., Leung, D.Y.M., 1992. Treatment of peanut allergy with rush immunotherapy. J. Allergy Clin. Immunol. 90, 256–262. doi:10.1016/0091-6749(92)90080-L

Patriarca, G., Nucera, E., Pollastrini, E., De Pasquale, T., Lombardo, C., Buonomo, A., Roncallo, C., Pecora, V., Musumeci, S., Altomonte, G., Alonzi, C., Schiavino, D., Gasbarrini, G., 2006. Oral Rush Desensitization in Peanut Allergy: A Case Report. Dig. Dis. Sci. 51, 471–473. doi:10.1007/s10620-006-3157-4

R Core Team, 2015. R: A Language and Environment for Statistical Computing.

Rimbaud, L., Heraud, F., La Vieille, S., Leblanc, J.C., Crepet, A., 2010. Quantitative risk assessment relating to adventitious presence of allergens in food: A probabilistic model applied to peanut in chocolate. Risk Anal. 30, 7–19. doi:10.1111/j.1539-6924.2009.01322.x

Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65. doi:10.1016/0377-0427(87)90125-7

Sørenson, T., 1948. A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons, Biologiske Skrifter // Det Kongelige Danske Videnskabernes Selskab. I kommission hos E. Munksgaard.

Spanjersberg, M.Q.I., Kruizinga, a. G., Rennen, M. a J., Houben, G.F., 2007. Risk assessment and food allergy: the probabilistic model applied to allergens. Food Chem. Toxicol. 45, 49–54. doi:10.1016/j.fct.2006.07.018

Taylor, S.L., Baumert, J.L., Kruizinga, A.G., Remington, B.C., Crevel, R.W.R., Brooke-Taylor, S., Allen, K.J., Houben, G., 2014. Establishment of Reference Doses for residues of allergenic foods: Report of the VITAL Expert Panel. Food Chem. Toxicol. 63, 9–17. doi:10.1016/j.fct.2013.10.032

van Rossum, C.T.M., Fransen, H., Verkaik-Kloosterman, J., Buurma-Rethans, E., Ocke, M., 2011. RIVM 2011, Dutch National Food Consumption Survey 2007-2010 : Diet of children and adults aged 7 to 69 years. RIVM-Rapport 350050006.

Wainstein, B.K., Studdert, J., Ziegler, M., Ziegler, J.B., 2010. Prediction of anaphylaxis during peanut food challenge: usefulness of the peanut skin prick test (SPT) and specific IgE level. Pediatr. Allergy Immunol. 21, 603–611. doi:10.1111/j.1399-3038.2010.01063.x

Zeiger, R.S., Sampson, H.A., Bock, S.A., Burks, A.W., Harden, K., Noone, S., Martin, D., Leung, S., Wilson, G., 1999. Soy allergy in infants and children with IgE-associated cow's milk allergy. J. Pediatr. 134, 614–622. doi:10.1016/S0022-3476(99)70249-0

3.7 Article 2: Combining food consumption data from different countries for creating food groups for allergen risk assessment (in Europe)

nutrients

MDPI

Article

Combining food consumption data from different countries for creating food groups for allergen risk assessment (in Europe)

Sophie BIROT [1,*], Charlotte B. Madsen [2], Astrid G Kruizinga [3], Amélie Crépet [4], Tue Christensen [2] and Per B. Brockhoff [1]

[1] DTU Compute, Richard Petersens Plads, DK-2800 Kgs. Lyngby, Denmark; sobi@dtu.dk
[2] National Food Institute, Technical University of Denmark, Denmark; charm@food.dtu.dk
[3] The Netherlands Organization for Applied Scientific Research (TNO), Zeist, The Netherlands; astrid.kruizinga@tno.nl
[4] ANSES, French Agency for Food, Environmental and Occupational Health Safety, 14 rue Pierre et Marie Curie, 94701 Maisons-Alfort, France; Amelie.CREPET@anses.fr
* Correspondence: sobi@dtu.dk; Tel.: +xx-xxx-xxx-xxxx

Received: date; Accepted: date; Published: date

Abstract: To prevent allergic reactions, food producers have to be able to make a knowledge based decision on whether to label their products with precautionary labelling. As many manufactured food products are also sold in different countries across Europe, the allergen risk assessment should be estimated at the European levels. As currently, there are no pan-European food data suitable for food allergy risk assessment. The aim of this paper is to investigate if intake data at a meal level from National Food consumptions Surveys can be combined to form a common European Food Intake database. We developed a procedure to investigate if national food intake can be combined and grouped using data from Netherlands, France and Denmark. The homogeneity of consumption patterns and the relevance of difference in risk of allergic reaction were assessed using a fixed framework of allergen concentration levels and threshold distribution. The groups formed were subsequently evaluated and adjusted based on practical considerations. It resulted in designing 62 food groups that can be used for allergen risk assessment. The summary statistics and descriptive names for each food group are included.

Keywords: Food allergy; National Food Consumption Surveys; food groups; probabilistic risk assessment

1. Introduction

Food allergy affects up to 6% of children and 3-4 % of adults and seems to be increasing [1]. This means that a large number of consumers must avoid consuming food products containing the harmful allergen. Checking the ingredients list is the first action to take to avoid the allergen. However, products with unintended allergen are potentially harmful for the allergic consumers and the unintended allergen may not be included in the ingredients list (European Directive 2003/89/EC). In order to prevent allergic reactions and inform consumers of the possible presence of allergens, precautionary labelling is used by food manufactures when a contamination is suspected. In EU, the use of precautionary labelling is not regulated. Therefore, this label is used on many products that do not contain the unintended allergen [2]. As a consequence of the lack of accuracy in labelling, allergic consumers may have dangerous behavior by ignoring precautionary labelling. To enable food producers to make a knowledge based decision on whether to label their products with precautionary labelling, it is necessary to perform a risk assessment. As many manufactured food

products are sold in different countries across Europe, a common way of estimating the risk of allergic reaction needs to be adopted. There are three elements in a food allergen risk assessment: data on the level of contamination with the allergen in the food, data from challenges studies with subjects with an allergy to the allergen, and data on how much of the food is normally consumed in a meal. As currently, there are no pan-European food data suitable for food allergen risk assessment, the aim is to investigate if consumption data from National Food Consumptions Surveys can be combined to form a common European food intake database at a meal level. In this initial step, we have developed a procedure to investigate if national food intake data can be combined using data from Netherland, France and Denmark. Data are combined in food groups with homogeneous consumption patterns. Aggregating food items eases the combination of national data having different level of detail and eases the choice of the user.

National Food Consumption Surveys are very detailed. Hence, selecting the correct food consumption is a condition for estimating the correct risk of allergic reaction in the population. In this paper, we describe the method developed based on the approach developed for within country consumptions' clustering and comparison [3]. The food groups resulting from the analyses are also illustrated with examples. A detailed description of all food groups including the accompanying consumption figures is also presented. The consumption data will be used in a risk assessment tool developed as a part of the EU project "Integrated Approaches to Food Allergen and Allergy Risk Management" (iFAAM).

2. Materials and Methods

The aim of the method was to design food groups across countries that could be used for allergen risk assessment. The different steps are described in Figure 1, and the following sections further explain it.

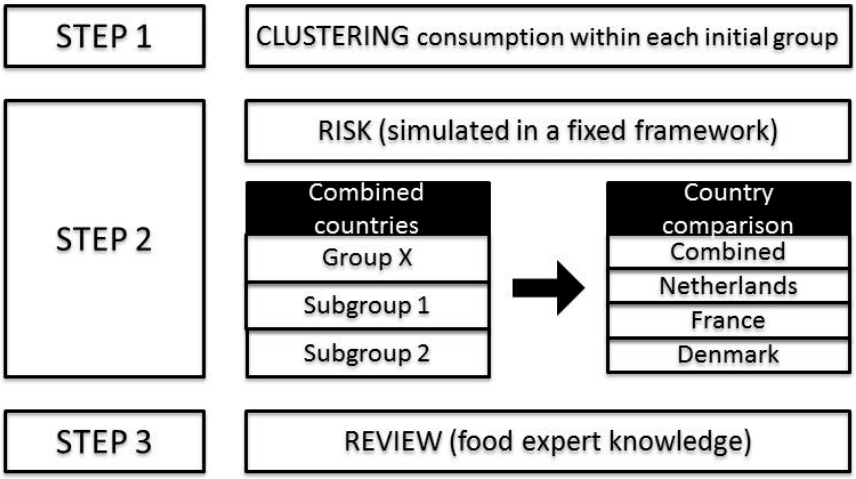


Figure 1: Overview of the steps of the method to design food groups across countries

2.1. National Food Consumption Surveys

National Food Consumption Surveys from Netherlands [1], France [2] and Denmark [3] were merged into one database to investigate the feasibility of creating a pan-European consumption database suitable for allergen risk assessment. The three National Food Consumption Surveys were combined using a common coding system for food items: The FoodEx 2 coding system developed by EFSA [4]. Participants are aged from 18 to 75 years. The food consumption was recorded for each eating occasion, defined as breakfast, lunch, dinner and other distinct eating occasions during the days.

2.2. Impact of the number of days in the surveys on the maximum consumption on an eating occasion

The relevant consumption scenario is the eating occasion. The reason is that the time between food intake and an allergic reaction is very low, so the allergen will not accumulate and therefore residuals allergen will not have impact on a later allergic reaction [5]. We chose a conservative approach and used the maximum consumption on a single eating occasion over the days for each food item consumed by each participant. The consumption surveys were not identical. The Netherlands performed two non-consecutive 24h recalls, whereas pre-coded 7 days food records were used in France and Denmark. However, it is expected that the maximum over 7 days is higher than the maximum over 2 days. As there are more chances to register higher consumption for a food item over 7 days than over 2 days. Thus, the influence on number of days on the maximum consumption distribution for each food item was first investigated for France and Denmark. In order to use the information available in the 7 days, two non-consecutive days were randomly selected among the 7 days in the survey. Then, the maximum consumption on an eating occasion is calculated over the two selected days per participants. Two non-consecutive days among the seven days in the surveys are sampled 50 different times, so the information available on the consumption of food items over the 7 days was used. Finally, the 50 maximum consumptions for all food items and all the consumers were averaged, so the numbers could be directly compared to the maximum consumption over 7 days. The relative difference in median is calculated to assess the difference between the original and the resampled data.

Furthermore, the impact of the number of days in the survey on the risk of allergic reaction was investigated. A relative difference in median from 5% to 35% was first simulated by shifting the consumption distribution with an incremental factor from 5% to 35%, the difference in risk was then compared to the risk calculated with the original data. The simulations' scenarios and its results are not detailed in this paper. It was found that if the risks' relative difference in median was 25% or higher, the difference in risk calculated could possibly be high. Consequently it was checked with the original and resampled data, that the impact on the risk difference was not too high, for the food item with relative difference in median higher than 25%.

2.3. Creating food groups for risk assessment

A procedure to design food groups suitable for allergen risk assessment was developed and described in [6].Groups with similar food items and similar consumptions were initially created by TNO based on consumption, portion size and if needed expert judgement. These groups were also validated using allergen risk assessment. Based on these initial groups, an adapted version was applied on the consumption data for the three countries combined. In short in the first step, a customized clustering method was applied to the combined consumption dataset, creating subgroups with homogeneous consumption patterns (step 1 on Figure 1). In the second step, a fixed framework was used to evaluate the impact of difference in consumption on the allergen risk assessment outcome. This framework used already published peanut and soy food challenged threshold distributions. A fixed set of concentration of unintended allergen (1, 10, 100 and 1000 ppm allergen protein) was also used to calculate the risk of allergic reaction patterns (step 2 on Figure 1). The combined distribution of food items consumptions within the food group and its food groups are used as inputs to the risk assessment to evaluate the impact of difference in consumption on the risk for groups of food items and their subgroups.

In the rest of the paper, "framework" is the set of concentration levels and distribution of allergens selected to support the evaluation of risk differences between a group and its subgroups. In this paper, the framework was used for two different purposes:

- First, the consequences of the differences between the maximum consumption over the 7 days or the 2 resampled days were assessed at the risk level.

- Secondly, the framework was used to assess the differences and similarities in risk, both, between the groups and subgroups designed during the clustering step, and between the countries.

The R software [7] was used to perform the simulation within the framework.

2.4. Maximum absolute difference and decision criterion

To evaluate how grouping and subgrouping influenced the risk between the groups a criterion to evaluate the difference in risk was developed in Birot et al., 2016. The selected criterion was the median of the maximum relative difference in risk calculated for the eight conditions within the framework (2 allergen x 4 concentration levels) for the combined countries (step 2 on Figure 1). This criterion was used in the same way for two purposes:

First, the criterion is used to decide whether an initial group was appropriate or should be divided into more groups (split). The maximum relative difference was calculated for the eights conditions of the framework in order to compare the subgroups risk to the overall groups risks.

$$max\ relative\ difference = \max\left(\frac{\left|Risk_{Split} - Risk_{\ Group}\right|}{Risk_{Group}}\right), \quad (1)$$

In a second step, the criterion was used to check if the combined countries consumption was representative enough or if a country specific consumption should be used to estimate the risk of allergic reaction for each designed group. The maximum relative difference was calculated for the eights conditions within the framework in order to compare the risk for the combined countries to the risk in Denmark, France and Netherlands.

$$max\ relative\ difference = \max\left(\frac{\left|Risk_{Country} - Risk_{\ Combined}\right|}{Risk_{Combined}}\right) \quad (2)$$

Finally, the criterion was used in the same way to compare the maximum relative differences to the criterion. If the eight relative differences were above the criterion, then the initial group was split or it was decided to use a country specific consumption.

2.5. Food group revision

New food groups based on the original groups were formed combining the consumption data from the three countries (Denmark, France and Netherlands) where possible. The last step was to name and describe the food groups, and to verify their validity by experts (step 3 on Figure 1). The purpose was indeed to have food groups that are logical and can easily be communicated to the end users of the food consumption data for food allergen risk assessment.

In some cases a group formed based on risk, could not be described in a logical way to the end user and the statistical analyses were overruled by practical considerations. This was done in the following way. In order to form a logical group we decided to use the consumption in the subgroup with the highest risk to characterize the food consumption in both groups.

2.6. Consumption's summary statistics per food groups

To describe the consumption of each food group, summary statistics were calculated from the food items' distributions within each food group defined after the three step procedure. Thus, food groups' distributions were defined combining the food items distribution within the food group of interest. Then, the summary statistics were calculated for consumers only. The mean, the standard deviation and percentiles are calculated per food group.

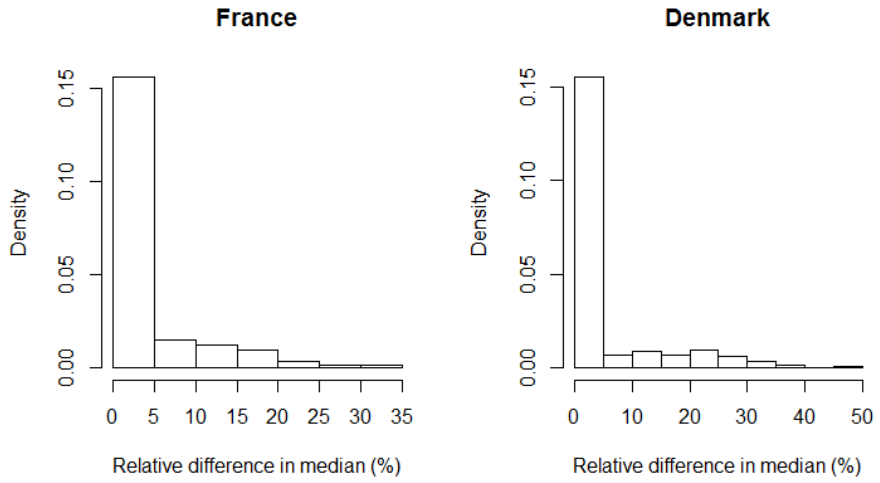
3. Results

3.1. Initial groups when combining National Food Consumption Surveys

The total number of participants was 8472 when the National Food Consumption surveys were merged (3819 in Netherlands, 2624 in France and 2029 in Denmark). The data included 703 different food items initially allocated to 42 food groups.

3.2. 2 vs 7 days comparison

Comparing the difference in maximum intake using 2- or 7-days intake showed that the relative differences in median between the original and the resampled distribution is small (Figure 2). The food item distributions for original and resampled surveys are quite closed for both France and Denmark. Figure 2 shows that most of the differences are lower than 5%. And as explained previously, until a 25% relative difference in median between the original and resampled consumption distributions, simulations scenarios has shown no high differences in risk. Therefore, it seems that the number of days in the survey does not have high impact on the food item consumption, hence the food grouping.



(**a**) (**b**)

Figure 2: Relative difference in median between the original and the resampled food items consumption distributions (**a**) in France; (**b**) in Denmark

The consequences on the risk assessment outcome were investigated only for the food items that were found to have a relative difference in median superior than 25%. Only 8 food items were found to have that high difference in France and 13 in Denmark. The average difference in risk for the food items with a high difference in risk is presented in Table 1 for the conditions within the fixed framework (2 allergens and 4 concentration levels). The difference in risk calculated with the original and the resampled consumption's distribution is small on average for France and Denmark. The risk for the resampled data is on average lower from 0 to 0.9 points in France and from 0.1 to 1.4 points in Denmark. The risks calculated with the original and the resampled data are presented in appendix A (France) and the appendix B (Denmark) for the food items with relative difference in median superior to 25%.

Allergen	Concentration levels							
	1 ppm		10 ppm		100 ppm		1000 ppm	
	France	Denmark	France	Denmark	France	Denmark	France	Denmark
Peanut	-0.1	-0.2	-0.3	-0.4	-0.5	-0.9	-0.9	-1.4
Soy	0.0	-0.1	-0.1	-0.1	-0.2	-0.3	-0.4	-0.6

Table 1: Average difference in risk between the original and the resampled consumption's distribution for the food items with relative difference in median superior than 25%

The risk calculated from the resampled consumption is lower because the maximum consumption over 2 days is lower than the one over 7 days. However, this difference is not high enough to have an impact on the risk outcome. Thus, the surveys' design available within the iFAAM project, 2 days in Netherlands vs. 7 days in France and Denmark, will have a limited impact on the group design.

3.3. Group subdivision based on clustering and risk assessment analysis

Each initial group was suggested to be divided in 1 or more subdivisions by the clustering method (Table 2).

Number of subdivisions suggested by the clustering	**No. of groups needing subdivision based on clustering**	**No. of groups needing subdivision based on risk assessment analysis**
0	5	25
1	29	13
2 or more	8	4

Table 2: number of subdivisions based on clustering and risk assessment analysis for the combined countries within each of the 42 initial groups

There were five initial groups for which no further split was suggested because they had a low number of food item (i.e. one or two), 29 with one subdivision and eight with two and more subdivisions. After the risk analysis, the number of initial groups that did not need subdivisions increased to 25. Simultaneously, the number of groups that needed one subdivision fell to 13 and the number of groups that needing two subdivisions fell to four.

3.4. Criterion for groups' design

The median of the maximum relative difference in risk was calculated across the conditions within the framework (2 allergens x 4 concentration levels) [6]. To illustrate the method, we can use the breakfast products' group as an example (Table 3).

Allergen	Concentration levels of contamination			
	1 ppm	10 ppm	100 ppm	1000 ppm
Peanut	41.7%	40.4%	35.9%	27.5%
Soy	43.5%	43.1%	42.1%	39.6%

Table 3: Maximum relative difference between the risk of the breakfast product group and its subgroups

In this case, the maximum difference for all the concentration levels and the 2 allergens are higher than the threshold, i.e. (around 12.6%). This group needs to be subdivided as the difference between the breakfast group's risk and its subgroups' risks is high according to the criterion.

3.5. *Initial groups with subdivisions*

Based on the criteria described above it was necessary to subdivide 40% of the 42 initial groups. This resulted in 17 food groups to be divided (Table 4). The number of subdivisions for each group that were found to be divided and their name is also indicated in Table 4.

Group name	**Number of subdivisions**
Chesnut paste and coconut milk	1
Milk powder and Cocoa powder	1
Milk(products), yoghurt(products), desserts	2
Potato chips	1
Meal replacements, meat imitates and supplements	1
Small sweets	1
Sugar	1
Sweet confectionary (jam, marmalade)	2
Sauces	1
Fish products	2
Meat products	1
Herbs and spices	1
Alcoholic drinks (excl. beer)	2
Cakes	1
Breakfast products	1
Fruit and vegetables, processed	1
Eggs and egg based dishes	1

Table 4: Groups with required a subdivision

The group division and the ones which were not divided resulted in 64 different food groups. The groups were renamed to match the actual content of the food groups newly designed.

Below we have used the breakfast group to illustrate the subdividing of a group. A large difference in risk (according to the criterion) can be seen between the two subgroups (Table 3 and Figure 3). It can be noted that the first subgroup has a higher risk than the second subgroup, so if the overall consumption will be used, it will results in an underestimation of the risk for the first subgroup and an overestimation for the second one. Moreover, the maximum relative differences between the risk of the breakfast products group and its subdivisions for each contamination level and each allergen are all above 12.6%.

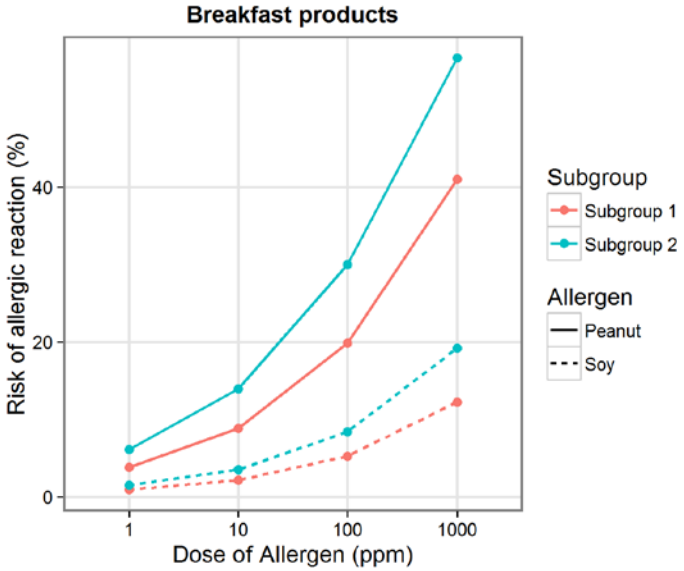


Figure 3: Risk of allergic reaction for two subgroups in the breakfast product group

3.6. Country comparison

The analyses resulted in 64 food groups. From these groups, only eight groups were country specific e.g. the consumption data from the three countries could not be merged based on the criteria chosen (Table 5).

Initial group name	Group name after subdivision	Subgroup
Fried/warm snacks	No change	No subdivision
Pancakes and waffles	No change	No subdivision
Chewing gum	No change	No subdivision
Sugar	No change	Subgroup 1
Sweet confectionary (jam, marmalade)	No Change	Subgroup 1
Meat products	Meat products - mean 65 g	Subgroup 1
Herbs/spices	Herbs and spices mixes, bouillon cubes, yeast extract	Subgroup 1
Breakfast products	Breakfast products, porridge	Subgroup 2

Table 5: Groups with country specific intake data

To illustrate the decision of having a country specific intake, the chewing-gum group is taken as an example. In Table 6, it can be seen that the maximum relative difference between risk, for each contamination level and allergens are all above the threshold of 12.6% (from 27.0% to 33.3%). It was therefore decided to create groups with country specific intake. In Figure 4, it can be noticed that the difference in risk between the three countries is large (according to the criterion).

Allergen	Concentration levels			
	1 ppm	10 ppm	100 ppm	1000 ppm
Peanut	29.9%	30.1%	28.8%	27.0%
Soy	33.3%	30.5%	31.1%	30.2%

Table 6: Maximum relative difference in risk between combined and individual countries for the chewing-gum group

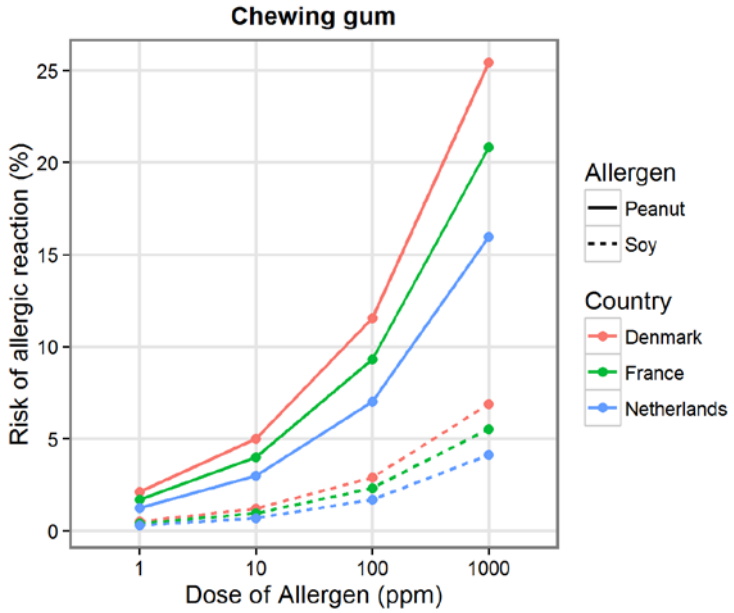


Figure 4: Risk of allergic reaction for the chewing-gum group in each country

3.7. Group for combined countries and without need of subdivision

Finally, 22 of the 42 initial groups were found to not require a subdivision and country combined consumption can be used. These food groups are shown in the Table 7.

Group name
Cheese
Cream and coffee milk
Ice cream
Peanuts, nuts and dried fruits
Soups
Chocolate and chocolate products
Cereal bars
Mashed potato powder
Potato product (excl. powder)
Vegetable oils and animal fat
Butter/halvarine/margarine
Crackers, crisp bread, rusk and toast
Bread, bread rolls and bread doughs
Beer
Syrups
Drinks without alcohol (excl. syrup)
Cookies (biscuits)
Binding agent
Pasta, rice, couscous and other grains
Legumes
Sandwich and pizza
Composite dishes

Table 7: Groups for combined countries and without need of subdivision

The cookies group illustrates the fact that the risk between the overall group and its subgroup is low, Figure 5. In Table 8, it can be seen that all maximum differences between the group risk and its subgroups is lower than the criterion of 12.6%. Furthermore, Table 9 shows that the maximum relative difference between the combined countries risk and each individual country risk is lower than the criterion. It was therefore decided that for the cookies group, the combined countries consumption is representative for all three countries.

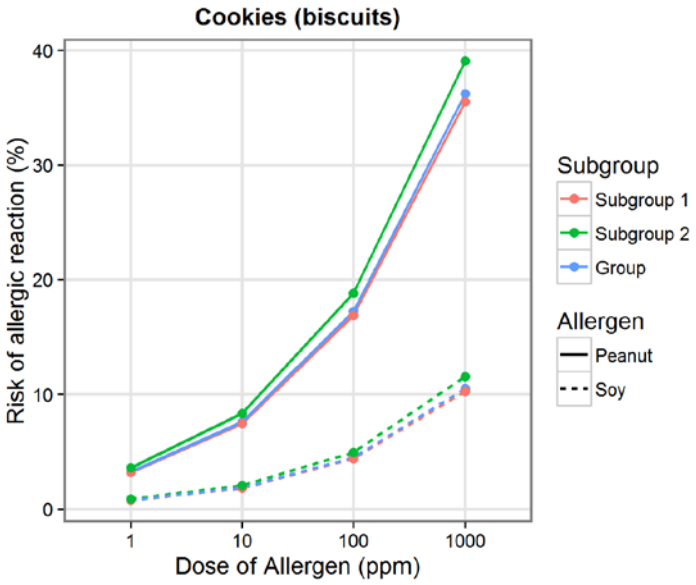


Figure 5: Risk of allergic reaction for the cookies group and its two subgroups

Allergen	Concentration levels			
	1 ppm	10 ppm	100 ppm	1000 ppm
Peanut	9.9%	10.0%	9.0%	8.0%
Soy	10.4%	10.3%	10.0%	9.8%

Table 8: Maximum relative difference in risk between the cookies group and its subgroups

Allergen	Concentration levels			
	1 ppm	10 ppm	100 ppm	1000 ppm
Peanut	10.9%	10.6%	9.2%	7.6%
Soy	11.2%	11.0%	10.6%	9.9%

Table 9: Maximum relative difference in risk between combined and individual countries for the cookies group

3.8. Adjustments based on practical considerations and groups' overestimation

The adjustments made for practical considerations for 12 food groups are summarized in Table 10. The reasons why some food groups were adjusted are detailed in Appendix C. As an example the sweet confectionary group was split into two groups with e.g. jam and 'other fruit spreads' in one subgroup and e.g. marmalade, rose hip jam and 'fruit spreads' in the other subgroup. In order to form a logical group, the consumption in the subgroup with the highest risk (jam) is used to characterise the food consumption in both groups. Furthermore, the group was described as such: Sweet confectionary (jam, marmalade). Similar, procedure was applied when the country with the highest risk was used to describe the food consumption of a food group.

Food group	Adjustments		
	Use subgroup with higher risk	Use country with higher risk	Use less detailed food item coding
Milk and milk products consumed with a spoon	✓		
Potato and other starch based chips (including salty sticks)	✓		
Sugar		✓	
Sweet confectionary (jam, marmalade)	✓		
Meat products – mean consumption 65 g		✓	
Herbs and spices mixes, bouillon cubes, yeast extract		✓	
Alcoholic drinks, alcohol ≤ 15%	✓		
Drinks without alcohol (excl. syrup)			✓
Cakes (including pastry)	✓		
Breakfast products, porridge		✓	
Fruit and vegetables, processed	✓		
Egg based dishes	✓		

Table 10: Summary of the adjustment made for practical consideration per food group

The relative difference in risk between a food group and the one used for the allergen risk assessment was average across the conditions within the framework (Table 11). For example, the consequences of using the French data for the "Meat 65g" consumption is that the Danish and Dutch consumption is overestimated with an average relative difference of 20% and 23%, respectively. It can be seen that using different food consumptions have different range of overestimation, from 6% to 74%. And on average, after adjustment for practical consideration, a food group will have a risk 27% higher than the one that would have been calculated with the original consumption data.

Group overestimated	Group name	Average relative difference between risks
Subgroup 3, Combined	Milk and milk products consumed with a spoon	47%
Subgroup 2, Combined	Potato and other starch based chips (including salty sticks)	74%
Subgroup 2, Combined	Sugar	24%
Subgroup 1, Netherlands	Sugar	28%
Subgroup 2, Combined	Sweet confectionary (jam, marmalade)	6%
Subgroup 3, Combined	Sweet confectionary (jam, marmalade)	19%
Subgroup 1, Denmark	Meat products – mean consumption 65 g	20%
Subgroup 1, Netherlands	Meat products – mean consumption 65 g	23%
Subgroup 1, France	Herbs and spices mixes, bouillon cubes, yeast extract	44%
Subgroup 1, Netherlands	Herbs and spices mixes, bouillon cubes, yeast extract	7%
Subgroup 1, Combined	Alcoholic drinks, alcohol ≤ 15%	8%
Subgroup 1, Combined	Cakes (including pastry)	22%
Subgroup 2, France	Breakfast products, porridge	29%
Subgroup 2, Netherlands	Breakfast products, porridge	14%
Subgroup 1, Combined	Fruit and vegetables, processed	31%
Subgroup 1, Combined	Egg based dishes	34%

Table 11: The relative difference in risk (overestimation) between a food group and the one used for the allergen risk assessment (see also appendix C for details of the groups)

A complete list of the 62 final groups created with this procedure, including consumption figures, can be found in Table 12.

4. Conclusion/discussion

In this paper, the first steps to create a pan-European consumption database, suitable for use in allergen risk assessment, were presented. The study showed that it is possible to merge consumption data and create common food groups relevant for food allergy risk assessment. The initial groups were adjusted for major inconsistencies and then based on the clustering method and the risk comparisons within the framework, the initial groups were split or validated without need of split. Furthermore, it was investigated if the consumption was representative enough or if a country specific consumption should be used to estimate the risk of allergic reaction. In a last step, some final adjustments were made for practical considerations.

The food manufacturers can export their food to many countries, so performing risk assessment for multiple countries is relevant. However, it was acknowledged that there is a lack of harmonized food consumption data across Europe [8]. Thus, in a purpose of creating data covering the whole Europe, the Food Consumption Surveys from Netherlands, France and Denmark were used together. Consumption data were recorded in a different way for the three countries. 24 hours recalls were performed over two days in Netherlands and the French and Danish food consumption was recorded with pre-coded seven days surveys. To comply with the tradition in food allergen risk assessment we chose a conservative approach using the maximum consumption on an eating occasion [5]. To our knowledge this is the first time consumption data obtained by the two survey methods have been compared using statistical methods. In order to insure the validity of using the maximum consumption over two and seven days within the same consumption database, a preliminary study was performed with the Danish and French data. Those surveys were selected because more information was recorded during the seven days of the surveys, which could fit in two

days designed. Kruizinga et al. (2008) proved using sensitivity analysis that the change in consumption has a low impact on the risk. Thus, as the differences in consumption between the original and the resampled consumption are small, it was expected that the difference in risk calculated with the original and resampled consumption to be also small. Furthermore, within the fixed framework selected to assess the impact of difference in consumption on the risk of allergic reaction, a slight overestimation of the risk calculated with the consumption over seven days was found even for the food items that were found to have high relative difference in median consumption (above 25%). So, it was concluded that those differences in surveys' design have a limited low impact on the final group designed.

The fixed framework of concentration levels and allergen threshold distribution presented in Birot et al. (2016) controlled the way the differences in consumption distribution were propagated to the risk. As the concentration levels and the threshold distribution are fixed within the framework, they have similar impact on the risk. Thus, the framework eases the comparison between the differences in consumption through the differences in risk. Furthermore, the group design is expected to be similar if using different condition in the fixed framework, as different concentration levels and different threshold distribution will have a proportional impact on the risk of allergic reaction.

As the amount of data from the three consumption surveys was very large, the method developed to create the food groups was required to be based on an automated criterion. Depending on the homogeneity of the initial group, the difference in risk between the overall group and its suggested subdivisions can be low or high. In order to evaluate the magnitude of the risk differences across food groups, a criterion was selected. The criterion choice was made, so no subjective information had influence. Based on some investigation (not detailed in this paper), the criterion was decided to be the median of the maximum relative difference in risk across food groups, this number is 12.6%. Thus, comparing the criterion within food groups makes sure the suggested subgroups with the clustering method actually detects a real difference in consumption distribution and not a difference in risk induced by the contamination level or allergen. It is also worth noting that the criterion is based on the probabilistic risk assessment using all data points of the combined distribution of food items within the food group and its food groups. Furthermore, the criterion also helps to assess the order of magnitude of risks differences across food groups and assure that differences between the group and its subgroups risks is high (according to the criterion). Thus, this criterion was found to be the best trade-off between homogeneity in consumption patterns within food groups and significant difference in risk across food groups.

To cope with the large dataset an automated approach was necessary, but it also resulted in groups that were not logical. A detailed assessment of the foods in the groups resulted in a manual adjustment of some of the groups in a way were the consumption in the combined new group was always based on the subgroup with the highest consumption. The disadvantage is that the risk of some products having a lower consumption will be overestimated. However, this principle ensures that consumption of a certain food will be overestimated rather than underestimated. The most extreme example is salty stickers were the risk will be 74% higher based on using the group data compared to the individual food. In average the manual approach for the twelve groups result in an overestimation of risk of 27%.

In hindsight adding salty stickers to the potato chip group, may not have been the wisest, but on the other hand creating a separate group for salty stickers also seems to be overkill. From appendix C it can be seen that the three food groups with an over estimation of more than 31% is; three products added to the "milk with a spoon" groups, one product added to "egg based dished" and the heterogeneous "Herbs and spices mixes, bouillon cubes, yeast extract" where the French consumption will be overestimated. For the other groups the overestimation of risk will be from 6 to 31%. We find that the excess risk is an acceptable price to pay for getting a relatively simple food grouping that is logical and can be communicated.

In this paper, we present an attempt to build the first pan-European consumption database suitable for allergen risk assessment based on food consumption data from Denmark, France and

363 The Netherlands. We have developed a system where food consumption survey data from different
364 countries on a meal level coded with FoodEx codes can be merged and logical food groups created.
365 To extend the database to more countries, it will be possible to organize similar coded food
366 consumption data from other countries into the defined food groups, and analyze similarities and
367 differences. Having intake data from across Europe and using the proposed procedures, it will be
368 possible to investigate, if one (or a few) pan-European food consumption data sets for use in food
369 allergy risk assessment can be created.

370 **Acknowledgments:** This work was part of the iFAAM project (Integrated Approaches to Food Allergen and
371 Allergy Risk Management, Grant Agreement No. 322147), the national food consumption data for France and
372 Denmark were kindly provided by the work package partners: ANSES and DTU. The national food
373 consumption data of the Netherlands was kindly provided by the National Institute for Public Health and the
374 Environment in the Netherlands. The challenge data for peanut and soy was kindly provided by TNO and
375 FARRP (University of Nebraska – Food Allergy Research and Resource Program).

376 **Author Contributions:** "Sophie BIROT, Charlotte B. Madsen, Astrid G Kruizinga, Amélie Crépet, Tue
377 Christensen and Per B. Brockhoff conceived and designed the experiments; Sophie Birot developed the
378 methods and analyzed the data with input from Charlotte B. Madsen, Astrid G Kruizinga, Amélie Crépet, Tue
379 Christensen and Per B. Brockhoff; Astrid G Kruizinga, Amélie Crépet, Tue Christensen and Per B. Brockhoff
380 contributed materials and analysis tools; Sophie Birot wrote the paper

381 **Conflicts of Interest:** The authors declare no conflict of interest. The founding sponsors had no role in the
382 design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in
383 the decision to publish the results.

384

385 **Table 12:** Food intake summary statistics per food group (in g)

iFAAM Name	Country	Mean	SD	P75	P90
Chesnut paste and coconut milk	Combined	63.7	55.3	100.0	116.5
Peanut butter	Combined	27.4	17.5	35.0	45.0
Cheese	Combined	39.8	31.3	48.0	80.0
Milk powder and Cocoa powder	Combined	18.5	14.1	26.4	33.6
Coffee creamer	Combined	4.3	4.2	6.0	8.0
Cream and coffee milk	Combined	22.5	26.1	30.0	40.0
Ice cream	Combined	88.2	47.3	100.0	150.0
Milk and milk products for drinking	Combined	264.5	163.7	317.5	432.0
Milk and milk products consumed with a spoon	Combined	156.9	76.4	200.0	250.0
Peanuts, nuts and dried fruits	Combined	33.3	29.5	40.0	60.0
Potato and other starch based chips (including salty sticks)	Combined	43.4	38.2	59.0	79.0
Fried/warm snacks	DK	162.8	103.1	180.0	270.0
Fried/warm snacks	FR	109.0	89.2	140.0	210.0
Fried/warm snacks	NL	77.4	50.5	85.5	140.0
Meal replacements and meat imitates	Combined	105.1	111.6	113.0	250.0
Supplements	Combined	1.7	2.6	2.0	3.0
Pancakes and waffles	DK	151.5	104.6	200.0	300.0
Pancakes and waffles	FR	152.7	102.3	200.0	300.0
Pancakes and waffles	NL	87.1	100.1	100.0	210.0
Soups	Combined	318.9	161.0	400.0	500.0
Small sweets - sweet confectionary unspecified/Combined	Combined	47.9	42.7	60.0	100.0
Small sweets - sweet confectionary specified	Combined	25.5	31.0	28.0	60.0
Sugar	Combined [1]	21.4	12.9	24.0	33.0
Chocolate and chocolate products	Combined	32.1	33.4	40.0	60.0
Sweet confectionary (jam, marmalade)	Combined	33.4	25.0	35.0	60.0
Cereal bars	Combined	31.7	27.1	32.1	50.0
Chewing gum	DK	10.6	7.8	10.0	20.0
Chewing gum	FR	5.9	7.4	6.0	10.0
Chewing gum	NL	2.9	2.5	4.0	5.0
Mashed potato powder	Combined	177.3	84.0	200.0	300.0
Potato product (excl. powder)	Combined	172.2	108.2	225.0	300.0
Vegetable oils and animal fat	Combined	14.8	11.9	20.0	30.0
Butter/halvarine/margarine	Combined	14.3	10.6	20.0	25.0
Sauces used as condiments and dessert sauces	Combined	22.3	20.7	30.0	46.5
Sauces , savory, chutneys and pickles	Combined	57.1	47.7	75.0	105.0
Fish products - 35 g such as fish fingers, fish paté	Combined	34.2	29.7	40.0	62.6
Fish products- mean 75 g such as smoked salmon, canned fish in oil	Combined	74.2	49.5	100.0	136.2
Fish products - mean 115 g such as fish cake, fish balls	Combined	115.6	75.0	150.0	190.0
Meat products - mean 65 g such as bacon, salami, paté	Combined [2]	64.5	46.2	75.0	125.0

Meat products -mean 105 g such as meat loaf, sausages	Combined	107.4	63.1	126.3	178.0
Crackers, crisp bread, rusk and toast	Combined	22.9	19.3	28.0	45.0
Bread, bread rolls and bread doughs	Combined	90.9	51.3	120.0	150.0
Herbs and spices mixes, bouillon cubes, yeast extract	Combined [1]	18.1	32.4	20.0	20.0
Spices and salt	Combined	2.9	2.5	3.0	4.0
Alcoholic drinks,alcohol ≤15%	Combined	222.1	144.5	282.5	420.0
Alcoholic drinks,alcohol above 15%	Combined	68.9	69.6	83.8	120.0
Beer	Combined	567.1	520.3	660.0	990.0
Syrups	Combined	26.6	30.8	33.8	60.4
Drinks without alcohol (excl. syrup)	Combined	362.3	252.1	483.3	600.0
Cookies (biscuits)	Combined	32.8	27.8	42.0	60.0
Cakes (including pastry)	Combined	144.4	78.0	180.0	250.0
Binding agent	Combined	10.7	14.5	13.0	26.9
Breakfast products eaten unprocessed (e.g. müsli, oat and maize flakes)	Combined	46.9	28.1	60.0	83.2
Breakfast products, porridge	Combined [1]	168.0	163.2	202.0	257.0
Pasta, rice, couscous and other grains	Combined	155.4	91.2	200.0	270.0
Legumes	Combined	132.2	67.3	175.0	215.0
Fruit and vegetables, processed	Combined	139.1	86.7	190.0	238.0
Eggs	Combined	40.9	29.0	55.0	80.0
Egg based dishes such as omelette	Combined	123.8	69.0	180.0	200.0
Sandwich and pizza	Combined	270.4	209.9	335.0	500.0
Composite dishes such as lasagna, quiche, vegetable casserole	Combined	238.2	155.5	320.0	450.0

386 [1] The group is combined but the consumption data used are the Danish (see also appendix C for
387 explanation)
388 [2] The group is combined but the consumption data used are the French (see also appendix C for
389 explanation)
390

391 **Appendix A:**

Allergen	Item	Name	Concentration levels							
			1 ppm		10 ppm		100 ppm		1000 ppm	
			Original	Resampled	Original	Resampled	Original	Resampled	Original	Resampled
Peanut	A004Q	Wheat germ	3.1%	2.9%	7.3%	6.9%	16.5%	15.7%	34.8%	33.3%
Peanut	A033J	Honey	2.8%	2.7%	6.5%	6.3%	14.9%	14.5%	31.9%	31.1%
Peanut	A03KA	Coffee beverages	6.4%	6.2%	14.7%	14.3%	31.5%	30.8%	59.0%	58.0%
Peanut	A03TJ	Soya drink	6.5%	6.3%	14.8%	14.3%	31.6%	30.8%	58.9%	58.0%
Peanut	A03YA	Veggie pot pie	5.9%	5.8%	13.5%	13.3%	29.4%	28.9%	56.4%	55.6%
Peanut	A043V	Savoury sauces	3.1%	3.0%	7.2%	7.1%	16.4%	16.1%	34.7%	34.2%
Peanut	A044V	Pesto	2.9%	2.9%	6.9%	6.8%	15.7%	15.6%	33.2%	32.9%
Peanut	A04NH	Fresh smoked sausages	5.2%	5.0%	12.0%	11.6%	26.3%	25.7%	51.8%	50.7%
Soy	A004Q	Wheat germ	0.7%	0.7%	1.8%	1.7%	4.3%	4.0%	10.1%	9.5%
Soy	A033J	Honey	0.7%	0.6%	1.6%	1.5%	3.8%	3.7%	9.0%	8.7%
Soy	A03KA	Coffee beverages	1.6%	1.5%	3.8%	3.6%	8.9%	8.6%	20.2%	19.7%
Soy	A03TJ	Soya drink	1.6%	1.5%	3.8%	3.7%	8.9%	8.7%	20.3%	19.7%
Soy	A03YA	Veggie pot pie	1.4%	1.4%	3.4%	3.3%	8.1%	8.0%	18.6%	18.3%
Soy	A043V	Savoury sauces	0.7%	0.7%	1.8%	1.7%	4.2%	4.1%	10.0%	9.8%
Soy	A044V	Pesto	0.7%	0.7%	1.7%	1.7%	4.0%	4.0%	9.5%	9.5%
Soy	A04NH	Fresh smoked sausages	1.2%	1.2%	3.0%	2.9%	7.1%	6.9%	16.5%	16.1%

392 **Table A1:** Difference in risk between the original and the resampled consumption's distribution,
393 France

394

Appendix B

Allergen	Item	Name	Concentration levels							
			1 ppm		10 ppm		100 ppm		1000 ppm	
			Original	Resampled	Original	Resampled	Original	Resampled	Original	Resampled
Peanut	A004Y	Wheat bread and rolls, white (refined flour)	4.8%	4.5%	11.2%	10.4%	24.7%	23.1%	49.3%	46.7%
Peanut	A023G	Cooked cured meat	3.3%	3.1%	7.6%	7.2%	17.3%	16.4%	36.5%	34.8%
Peanut	A025J	Cooked sausages	3.3%	3.2%	7.7%	7.4%	17.6%	16.8%	36.9%	35.6%
Peanut	A02MA	Cow milk, skimmed (low fat)	7.9%	7.5%	17.8%	17.1%	37.4%	36.1%	67.1%	65.6%
Peanut	A032J	White sugar	2.3%	2.2%	5.5%	5.2%	12.7%	12.1%	27.7%	26.5%
Peanut	A034V	Sweet confectionery	3.8%	3.6%	8.9%	8.4%	19.9%	19.0%	40.8%	39.3%
Peanut	A034X	Candies	2.8%	2.6%	6.5%	6.1%	14.7%	13.9%	31.5%	29.9%
Peanut	A03KC	Coffee (average strenght) beverage	9.2%	8.7%	20.7%	19.7%	42.6%	40.8%	73.2%	71.3%
Peanut	A03ND	Cider	7.6%	7.4%	17.2%	16.9%	36.3%	35.9%	65.9%	65.4%
Peanut	A03VN	Hummus	3.0%	3.0%	7.1%	7.0%	16.3%	15.9%	34.6%	34.0%
Peanut	A042D	Mixed vegetable salad	4.6%	4.3%	10.6%	10.1%	23.3%	22.4%	46.1%	44.6%
Peanut	A044F	Table-top condiments	3.1%	3.0%	7.3%	7.1%	16.6%	16.2%	35.3%	34.3%
Peanut	A044X	Mayonnaise	2.4%	2.3%	5.6%	5.4%	13.0%	12.5%	28.4%	27.3%
Soy	A004Y	Wheat bread and rolls, white (refined flour)	1.2%	1.1%	2.8%	2.6%	6.7%	6.1%	15.5%	14.3%
Soy	A023G	Cooked cured meat	0.8%	0.7%	1.9%	1.7%	4.5%	4.2%	10.6%	10.0%
Soy	A025J	Cooked sausages	0.8%	0.7%	1.9%	1.8%	4.5%	4.3%	10.7%	10.2%
Soy	A02MA	Cow milk, skimmed (low fat)	1.9%	1.8%	4.6%	4.4%	10.9%	10.4%	24.4%	23.4%
Soy	A032J	White sugar	0.5%	0.5%	1.3%	1.3%	3.2%	3.0%	7.6%	7.2%
Soy	A034V	Sweet confectionery	0.9%	0.9%	2.2%	2.1%	5.2%	5.0%	12.3%	11.7%
Soy	A034X	Candies	0.6%	0.6%	1.6%	1.5%	3.8%	3.5%	9.0%	8.4%
Soy	A03KC	Coffee (average strenght) beverage	2.3%	2.1%	5.5%	5.1%	12.8%	12.1%	28.3%	26.9%
Soy	A03ND	Cider	1.8%	1.8%	4.4%	4.4%	10.5%	10.3%	23.6%	23.3%
Soy	A03VN	Hummus	0.7%	0.7%	1.7%	1.7%	4.2%	4.1%	9.9%	9.7%
Soy	A042D	Mixed vegetable salad	1.1%	1.0%	2.6%	2.5%	6.3%	6.0%	14.6%	13.9%
Soy	A044F	Table-top condiments	0.7%	0.7%	1.8%	1.7%	4.3%	4.2%	10.1%	9.8%
Soy	A044X	Mayonnaise	0.6%	0.5%	1.4%	1.3%	3.3%	3.1%	7.8%	7.5%

Table A2: Difference in risk between the original and the resampled consumption's distribution, Denmark

Appendix C: Detailed adjustments made for practical considerations

Milk and milk products consumed with a spoon

Three foods, that had their own subgroup, but fit the description of this group and had a lower risk, were moved into this group without including them in the combined consumption of the group.

Potato and other starch based chips (including salty sticks)

Salty sticks and Tapioca starch-based snacks with a lower consumption were moved into the group without contributing to the group consumption. This means that the risk connected to these two foods will be overestimated.

Sugar

The risk analyses divided the sugar group in two groups both including white sugar. To solve this one group was formed using the Danish sugar consumption data to characterise the group. This means that the FR and NL sugar intake will be overestimated.

Sweet confectionary (jam, marmalade)

This group was split in two groups with e.g. jam and 'other fruit spreads' in one subgroup and e.g. marmalade, rose hip jam and 'fruit spreads' in the other subgroup. In order to form a logical group the two subgroups were combined into one group: Sweet confectionary (jam, marmalade). The consumption in the subgroup with the highest risk was used to characterise the food consumption in the combined group.

Meat products – mean consumption 65 g

To be able to form two logical groups of the meat consumption data the FR data was used to characterise this group. This means that the DK and NL risk for this food group will be overestimated. In the other meat group the combined data from the three countries was used.

Herbs and spices mixes, bouillon cubes, yeast extract

This is a small and heterogeneous group. To make a combined group the Danish consumption data was used: This means that the FR and NL risk for this group will be overestimated.

Alcoholic drinks, alcohol ≤ 15%

The risk analyses resulted in two groups with wine in one group and red wine in the other (different levels of reporting consumption). The two groups were merged, but the consumption used is the one representing the subgroup with the highest risk.

Drinks without alcohol (excl. syrup)

A less detailed level of coding was used due to a very detailed food items coding in this group. The very detailed coding made it impossible to form meaningful group. Instead of the original 73 food items, the food items were grouped based on a higher level of coding. The higher level of coding organised the 73 food items in seven levels:

- Coffee, cocoa, tea and herbal drinks
- Coffee, cocoa, tea and herbal ingredient
- Drinking water
- Fruit juices and nectars
- Other fruit and vegetable juices or nectars
- Vegetable juices

This group did not need to be subdivided.

Cakes (including pastry)

The risk analyses resulted in two groups with many cakes but no logical division e.g. 'croissant' in one group and 'croissant filled with cream' in the other. The two groups were merged, but the consumption used is the one representing the subgroup with the highest risk.

Breakfast products, porridge

This is a small group. To make a combined group the Danish consumption data was used: This means that the FR and NL risk for this group will be overestimated.

Fruit and vegetables, processed

This is a large and heterogeneous group. The risk analyses resulted in two groups with no logical division: e.g. 'fruit compote' in one group and 'fruit compote, apple' in the other. The two groups were merged, but the consumption data used for the group is the one representing the subgroup with the highest risk.

Egg based dishes

The risk analyses resulted in two groups. One food item was moved from the group with the lower risk to the group with the higher risk without changing the consumption data in order to make two groups that could be described logical.

References

1. van Rossum, C. T. M.; Fransen, H.; Verkaik-Kloosterman, J.; Buurma-Rethans, E.; Ocke, M. *RIVM 2011*; 2011.

2. Dubuisson, C.; Lioret, S.; Touvier, M.; Dufour, A.; Calamassi-Tran, G.; Volatier, J.-L.; Lafay, L. Trends in food and nutritional intakes of French adults from 1999 to 2007: results from the INCA surveys. *Br. J. Nutr.* **2010**, *103*, 1035–48.

3. Pedersen, A. N.; Fagt, S.; Groth, M. V.; Christensen, T.; Biltoft-Jensen, A.; Matthiessen, J.; Andersen, N. L.; Kørup, K.; Hartkopp, H.; Ygil, K. H.; Hinsch, H.-J.; Saxholt, E.; Trolle, E. *Danskernes kostvaner 2003-2008*; 2008.

4. European Food Safety Authority The food classification and description system FoodEx 2 (revision 2). *EFSA Support. Publ.* **2015**, *8*.

5. Madsen, C. B.; Houben, G.; Hattersley, S.; Crevel, R. W. R.; Remington, B. C.; Baumert, J. L. *Risk Management for Food Allergy*; Elsevier, 2014.

6. Birot, S.; Madsen, C. B.; Kruizinga, A. G.; Christensen, T.; Crépet, A.; Brockhoff, P. B. Grouping food consumption data for use in food allergen risk assessment. *Manuscript submitted for publication* **2016**.

7. R Core Team R: A Language and Environment for Statistical Computing 2015.

8. Brussaard, J. H.; Lowik, M. R. H.; Steingrimsdottir, L.; Moller, A.; Kearney, J.; De Henauw, S.; Becker, W.; grp, E. A European food consumption survey method - conclusions and recommendations. *Eur. J. Clin. Nutr.* **2002**, *56*, S89–S94.

9. Kruizinga, A. G.; Briggs, D.; Crevel, R. W. R.; Knulst, A. C.; Bosch, L. M. C. van den; Houben, G. F. Probabilistic risk assessment model for allergens in food: sensitivity analysis of the minimum eliciting dose and food consumption. *Food Chem. Toxicol.* **2008**, *46*, 1437–1443.

Chapter 4

Probabilistic risk assessment

4.1 Aim and outline of the chapter

As explain in chapter 1, one of the aim of the PhD project was to review the different methods used to estimate the risk of allergic reaction. Two already published methods were developed by two project partners. A simulation method based on first order Monte-Carlo simulations was developed by TNO [14]. And a method based on second order Monte-Carlo simulations and Bayesian inferences was developed by ANSES [13]. The results of the investigations are presented in the article 3: "Allergen probabilistic risk assessment modelling: existing model comparison and proposition of an alternative Frequentist approach to account for uncertainty". This article is a draft version and will be submitted when co-authors will give feedback.

First, an example using three log-normal distribution to fit the distribution, contamination and threshold distributions help to understand mathematical the risk simulation and uncertainty propagations. Then, the risk is estimated for the different methods investigated in the paper: the already two existing method and a third method based on second order Monte-Carlo simulations. Comparisons on the risk estimation with the methods are made. The uncertainty propagation is evaluated with the different methods. So, it can understand from which distri-

bution the uncertainty on the risk is coming from. Thus, it will to identify the distribution for which the data collection should be improved, given the data used in the article.

In appendix A, some investigations on fitting the survival model to the threshold data with the Survival package in R and using Bayesian simulations with the JAGS software are reported.

4.2 Article 3: Allergen probabilistic risk assessment modelling: existing model comparison and proposition of an alternative Frequentist approach to account for uncertainty

This paper is a draft paper and further work will be made on it before publication.

Allergen probabilistic risk assessment modelling: existing model comparison and proposition of an alternative frequentist approach to account for uncertainty

Contents

1. INTRODUCTION

Food allergy had been a growing public health concern over the last decade with around 3-5% of the adults and 8% of the children that suffer from allergic reaction (1). Allergic persons need to avoid consuming food products containing the offending allergen (2) to prevent an unexpected allergic reaction. To support this, the ingredient labelling should provide essential information to allergic individuals on which allergens are present in the food product as an ingredient. However, unexpected allergic reactions can still occur due to the unintended presence of allergen in the food products which are not in the list of ingredients. In order to warn allergic consumers and avoid dangerous allergic reaction, food manufacturers use “may contain” labelling when cross contamination was suspected or facilities are shared (3).

The increasing use of “may contain” labelling has lead allergic persons to disregard the warning, as food choice can be drastically reduced (4). In order to help manufacturers use “may contain” labelling based on a risk assessment focussing on the health risk for the consumer, it was recommended to use quantitative risk assessment (5). Thus, the consequences of unintended presence of allergen in food products can be quantified at a population level and a knowledge based recommendation can be made to food manufacturers to apply labelling.

ANSES and TNO both developed probabilistic risk assessment for food allergy based on different approaches (6,7). The two approaches have the same three input model variables: the consumption distribution (i.e. how much of the suspected contaminated product is consumed), the concentration distribution (i.e. how much allergen is in the contaminated product) and the threshold distribution (i.e. how much allergen do the allergic consumers react to). However, one modelling method is only based on first order Monte Carlo simulations (7) and the other one is based on the combination of Bayesian inferences and Monte-Carlo simulations (6). These methods both take into account the variability (i.e.

heterogeneity among the population) and the uncertainty (i.e. lack of knowledge about the true value). The last one based on second order Monte-Carlo simulations make possible to separate the variability and the uncertainty. The purpose of this paper is to mathematically compare the two methods and to understand the mechanism of uncertainty propagation from the inputs parameters to the risk. The comparison will be done using a simplify model based on log-normal distributions for the three input variables. A focus will be done on how the uncertainty propagates, and how the risk can be expressed. This simple mathematical expression will make possible to mathematically calculate the uncertainty on the risk and to compare with results from simulations. Based on this proposed model, extension of the risk estimation model will be proposed and compared to the existing approaches. Thus, the two existing models had been adapted to estimate the risk of allergic reaction: the risk for the allergic consumers when all the products were contaminated. A study case of contamination of cereal bars with peanut allergen will illustrate the model comparisons.

2. MATERIELS AND METHODS

2.1. General mathematical formulation

The two approaches developed by ANSES and TNO have the same input variables such as consumption, concentration and threshold distribution The mathematical formulation of both models can be generalize as follow.

The general mathematical formulation of the model is formulated, so the similarities and differences between the various ways of modelling can be explicitly clarified and the consequences on the risk assessment can be formally assessed. The consumption X follows a distribution noted F_X, the concentration Y follows a distribution called F_Y and the threshold Z follows a distribution called F_Z. The allergy outcome U is defined as follow:

$$\begin{cases} U=1, \text{ if } Z < XY, \text{ and} \\ U=0 \text{ otherwise} \end{cases}$$

U follows a Bernoulli distribution with probability p_u, where $p_u = P(Z < XY)$. Formally it is assumed that the three random variables X, Y and Z are independent. The probability p_u is mathematically a function of the three distributions: $p_u = f(F_x, F_y, F_z)$. The general principle of the risk modelling is presented on Figure 1.

If these three distributions are given by parameters, then this a function of these:

$$p_u = f(\theta_x, \theta_y, \theta_z)$$

Furthermore, the uncertainty on the parameters distribution propagates to the risk estimation, the differences in the way the uncertainty propagates with the two methods will also be assessed. To illustrate the way the risk is calculated with the two methods, an example inspired from (6) is used. The input variables distributions are described, so the difference in risk simulation can be assessed.

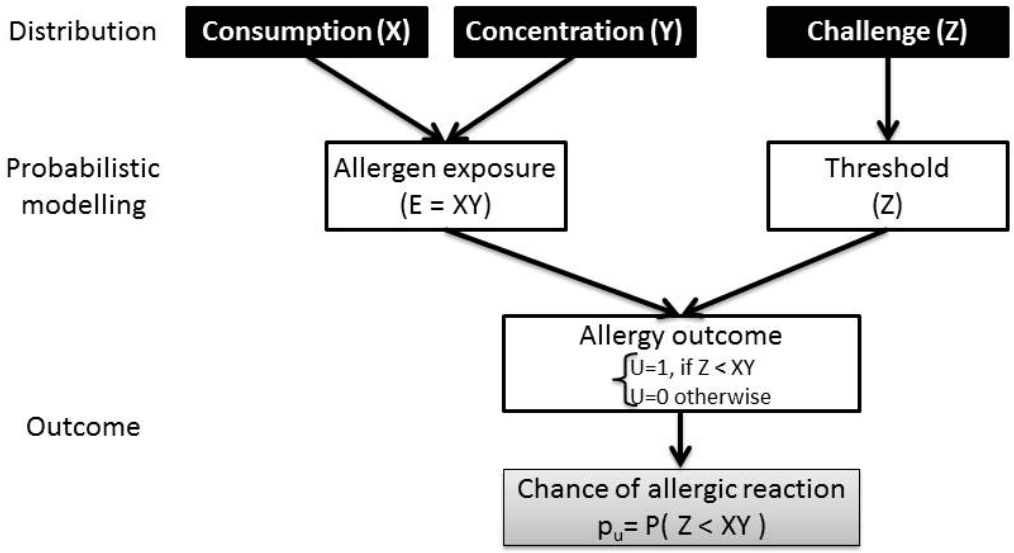


FIGURE 1: RISK ESTIMATION – GENERAL PRINCIPLE

2.2. Mathematical formulation of the model developed by TNO

The probabilistic risk assessment method described in (7) aims to take into consideration the variability and the uncertainty from input variables. The risk for the allergic user population can be expressed with a mathematical formulation, the three input distributions are stated in the article as:

X_{ik} ~ Log-Normal(μ_X,σ_X) (Consumption distribution)

Y_{ik} ~ Log-Normal(μ_Y,σ_y) (Concentration distribution)

Z_{ik} ~ empirical distribution of threshold

where *i* is the iteration and *k* the run.

The simulations are repeated for *n* (1000) iterations and *K* (25) runs from which the risks uncertainty and variability are calculated.

For each (i,k), X_{ik}, Y_{ik} and Z_{ik} are simulated from the distributions, if $X_{ik}Y_{ik} > Z_{ik}$, then $U_{ik}=1$, otherwise $U_{ik}=0$. Thus, the risk is calculated for each run *k*:

$$R_k = \frac{1}{n}\sum_{i=1}^{n} U_{ik}$$

And the standard deviation is calculated for the K risks:

$$sd = \sqrt{\sum_{k=1}^{K} \frac{(R_k - \bar{R})^2}{K-1}}$$

An attempt to calculate the risk's uncertainty and variability was made by adding uncertainty on the scale parameters (σ) of the Log-Normal consumption and concentration distributions and by adding variability within persons for the threshold distribution. However, it can be noted that the parameters

do not vary from run to run (k). So, uncertainty and variability introduced do not actually contribute to estimate the risk's uncertainty. And, increasing the number of simulations, i.e. increase the number of iterations or/and runs, will lead to calculate the risk with a better accuracy and as the standard deviation calculated by this method express the uncertainty of the numerical procedure and not the uncertainty and variability introduce by inputs variables.

2.3. Mathematical formulation of the model developed by ANSES

The probabilistic risk assessment described in (6) uses a combination of second order Monte-Carlo simulations and Bayesian inferences to estimate the risk of allergic reaction. The three input distributions described in the previous sections can be used to express mathematically the risk for the allergic user population. In Bayesian analysis the parameters distribution are also characterized with distributions. Thus, the actual input distributions and their prior parameters' distribution are defined below:

X_{ik} ~ empirical distribution with sampling with replacement with survey's weights

Y_{ik} ~ Exponential(λ_{Yk}) with prior distribution: λ ~ Gamma(10^{-3},10^{-3})

Zik ~ Weibull(a_{Zk},b_{Zk}) with prior distribution: (a,b) ~ Gamma(10^{-3},10^{-3})

The parameters' posterior distributions are estimated either by direct calculation when prior distribution are conjugate or using second order Monte-Carlo simulations. It resulted in parameters distributions from which K (=100) set of parameters are sampled in order to integrate the uncertainty in the risk estimation. n (=number of participants in the Food Consumption Survey) iterations are made for each set of parameters. Then, the risk is estimated at the individual level from the dose response curve:

$$Risk_{ik} = \text{Dose-Response}(Exposure_{ik}) \text{ with } Exposure_{ik} = X_{ik} \times Y_{ik}$$

The risk represents the probability that the allergic consumer will react to the amount of peanut protein ingested. Thus, distributions of risks are obtained to describe the uncertainty and variability in risks.

2.4. Data description

2.4.1.Concentration of unattended allergen distribution (Y)

In order to be able to compare the two methods for the risk modelling, peanut concentration data in cereal and nutrition bars were collected in publications (8). Thus, the 24 data points collected were used as an input to the risk assessment.

2.4.2.Consumption distribution (X)

The cereal bars combined consumption in Netherlands, France and Denmark is used as an input distribution to the risk assessment (9). As allergic reactions are acute consumption on a single eating occasion is used. In addition a conservative approach is used, using the maximum consumption for each consumer.

2.4.3.Challenge threshold distribution (Z)

Doses that elicit an objective allergic reaction are collected in double blind placebo controlled food challenge (DBPCFC) with peanut protein. Survival models can be fitted to the threshold points (10). The NOAEL (No Observable Adverse Effect Level) and LOAEL (Lowest Observable Adverse Effect Level) are the discrete values with which the survival is fitted. Thus, an interval-censoring survival analysis was found to be the appropriate model to analyze such data (11), as the exact dose that causes the allergic reaction is not known but falls between the NOAEL and the LOAEL. Weibull, Log-logistic and Log-Normal cumulative probability function are usually used to model individuals threshold of allergic reaction (11). The dose response (DR) curve is used to describe the link between the proportions of allergic consumers reacting to a given amount of peanut protein (exposure to the allergen: E). Thus, for the Log-normal distribution the relationships with the exposure (E) can be expressed as:

$\mathrm{DR(E)} = \phi(\frac{\ln(\mathrm{E}) - \mu_Z}{\sigma_Z})$ where ϕ is the cumulative distribution function of the standard normal distribution,

and for the Weibull distribution: $\mathrm{DR\ (E)} = 1 - \ \mathrm{e}^{-\left(\frac{\mathrm{E}}{b_Z}\right)^{a_Z}}$.

A selection of the data described in (10) was used to describe the response to peanut DBPCFCs. Thus, the distributions parameters were estimated with 158 NOAEL and LOAEL values collected in publication (12–23).

2.5. Methods comparison and suggestion of an alternative model

As described in sections 2.2 and 2.3 the two models use the input distribution in different ways to estimate the risk of allergic reaction. After mathematical review, it was highlighted that the method in (7), even if it claims to, does not actually integrate the risk uncertainty. However, the method described in (6) makes explicit the uncertainty propagation from the input variables to the risk calculation.

Besides the risk's uncertainty, the two methods also differ from the way the threshold model is used. In (7), a threshold is simulated for the distribution for each individuals and then compared to the amount of allergen ingested. The risk is then calculated by counting the number of reaction among the simulated consumers. Whereas in (6), the dose response curve is used to predict the chance of allergic reaction for each consumer. Distributions of users risk are used to estimate the risk of allergic reaction and its uncertainty. As the same threshold distribution is used in both methods for this study (i.e. not in the original papers), the difference in use of this distribution should not impact the risk estimation. It will be illustrated in the result part with the risk estimation.

An alternative method that includes uncertainty in the risk calculation and that is not based on Bayesian inferences is proposed. In order to is to see if a non-Bayesian approach is able to also take into account uncertainty. Inspired from (7), Monte-Carlo simulations that include parameters sampling can be used.

The parameters distributions are defined from well-known distributions. The detailed distribution from which the parameters are sampled will be detailed in the next sections.

2.6. Uncertainty analysis (uncertainty propagation)

2.6.1. Uncertainty analysis – general principal

An adapted version of the uncertainty analysis described in (24) to evaluate sources of uncertainty individually to match the recommendation made by EFSA (25) was used to assess how the uncertainty propagates from the consumption, contamination and threshold distributions separately to the risk. Thus, it can be identified which distribution add the most uncertainty to the risk.

It was investgated how the sampling uncertainty of all the distributions propagates through the non-linear $\hat{p}_u$ computation. This can be investigated actually for each distribution separately of for all of them jointly. Formally, it amounts to considering the $\hat{p}_u$ as a random variable as a function of the data:

$$\hat{P}_u^X = f(\widehat{\Theta}_X, \hat{\theta}_y, \hat{\theta}_z)$$

where $\widehat{\Theta}_X$ is a random sampling statistics which investigated the uncertainty induced by consumption data sampling, and the other two ones are held at the observed estimated parameters. Or, to investigate the uncertainty induced by the concentration data sampling:

$$\hat{P}_u^Y = f(\hat{\theta}_x, \widehat{\Theta}_Y, \hat{\theta}_z)$$

Or, to investigate the uncertainty induced by the threshold data sampling:

$$\hat{P}_u^Z = f(\hat{\theta}_x, \hat{\theta}_y, \widehat{\Theta}_Z)$$

Or all of them, which is how the risk of allergic reaction is usually calculated whith the probabilisitc modelling. The uncertainty induced by the three distribution is investigated:

$$\hat{P}_u = f(\hat{\Theta}_X, \hat{\Theta}_Y, \hat{\Theta}_Z)$$

2.6.2.Triple-log-normal uncertainty propagation mathematical formulation

When three log-normal distributions are used to calculate the risk of allergic reaction, the probability can be expressed analytically and the the risk is function of 6 parameters:

$$p_u = f\left(\mu_x^L, \sigma_x^L, \mu_y^L, \sigma_y^L, \mu_z^L, \sigma_z^L\right)$$

In fact, in this case it becomes:

$$p_u = 1 - \Phi\left(\frac{\mu_z^L - \mu_x^L - \mu_y^L}{\sqrt{(\sigma_x^L)^2 + (\sigma_y^L)^2 + (\sigma_z^L)^2}}\right)$$

Thus, using three log-normal distributions, besides estimating the risk for an allergic reaction for allergic users, has the advantage that the risk distribution can be expressed in a mathematical simple expression. It will help to assess the uncertainty input parameter propagation to the risk, but it is the recommended that we always use the log-normal distribution for all the input distributions. The risk's uncertainty can indeed be calculated with the multivariate propagation of error formula as the risk is estimated with three independent distributions (26). The uncertainties should approximately add up on the variance scale in the triple log-normal model:

$$Var(p_U) \approx (\frac{\partial p_u}{\partial \mu_x^L})^2 \sigma^2_{\mu_x^L} + (\frac{\partial p_u}{\partial \mu_y^L})^2 \sigma^2_{\mu_y^L} + (\frac{\partial p_u}{\partial \mu_z^L})^2 \sigma^2_{\mu_z^L} + (\frac{\partial p_u}{\partial (\sigma_x^L)^2})^2 \sigma^2_{(\sigma_x^L)^2} + (\frac{\partial p_u}{\partial (\sigma_y^L)^2})^2 \sigma^2_{(\sigma_y^L)^2} + (\frac{\partial p_u}{\partial (\sigma_z^L)^2})^2 \sigma^2_{(\sigma_z^L)^2}$$

The derivative calculation is detailed in appendix 1, and the variance of the risk was found to be:

$$Var(p_U) \approx \sum_{i=(X,Y,Z)} K_1 \frac{\sigma_i^2}{n_i} + \sum_{i=(X,Y,Z)} K_2 \frac{2\sigma_i^4}{(n_i - 1)}$$

where n_i is the number of data points for each input data, and

$$K_1 = \left[\phi\left(\frac{\mu_z^L - \mu_x^L - \mu_y^L}{\sqrt{(\sigma_x^L)^2 + (\sigma_y^L)^2 + (\sigma_z^L)^2}} \right) \cdot \frac{1}{\sqrt{(\sigma_x^L)^2 + (\sigma_y^L)^2 + (\sigma_z^L)^2}} \right]^2$$

and

$$K_2 = \left[\phi\left(\frac{\mu_z^L - \mu_x^L - \mu_y^L}{\sqrt{(\sigma_x^L)^2 + (\sigma_y^L)^2 + (\sigma_z^L)^2}} \right) \cdot \frac{\mu_z^L - \mu_x^L - \mu_y^L}{2\left(\sqrt{(\sigma_x^L)^2 + (\sigma_y^L)^2 + (\sigma_z^L)^2}\right)^3} \right]^2$$

Thus, the risk uncertainty can be calculated mathematically in the triple log-normal case, for each source of uncertainty separately and the three combined. So, it can be actually checked that the uncertainty from the three sources add up.

2.7. Uncertainty propagation – sampling scheme

2.7.1.General principle

The K different sets of parameters are sampled for the three distributions with parameters calculated with the examples presented in section 2.1. And n simulations are performed with the K different set of parameters. In order to assess the magnitude of the uncertainty from the three distributions individually, as explained in Figure 2, one distribution of parameters is selected for one input distribution at a time. Then, the risk is calculated given the uncertainty from one input distribution only. This step is repeated for each input distribution, so the uncertainty on the risk coming for each input can be estimated individually. As the inputs are assumed to be independent, the sum of the three uncertainties is compared to the uncertainty on the risk calculated with the uncertainty on the three input distribution at the same time. The risk was estimated in four different ways, so the comparison between the different ways of simulated the risk can be done. The different methods were described in sections 2.2, 2.3 and 2.4. In each case, the uncertainty on the parameters' input distribution is integrated for one input distribution at a time. So the sum of uncertainties can be compared to the

uncertainty when the uncertainties of all parameters' inputs distribution are integrated at the same time.

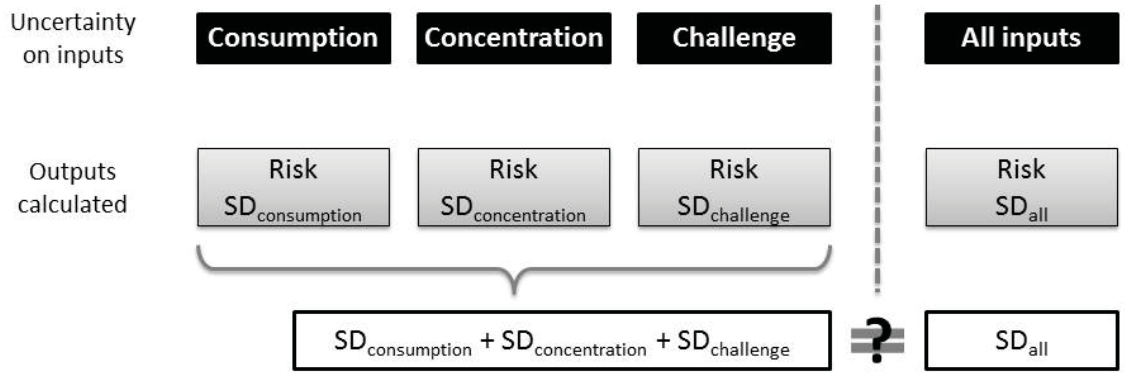


FIGURE 2: UNCERTAINTY ON THE PARAMETERS SAMPLING SCHEME (SD= STANDARD DEVIATION)

2.7.2.Sampling distributions for the triple log-normal case

In the case three log-normal distributions are used for the estimation of the risk of allergic reaction, the probability distributions for the parameters were sampled from well-known distributions. For each distribution *i* with n_i data points (consumption, contamination and challenge), the two parameters μ and σ of the log-normal distribution are sampled from:

- μ_i: the mean is sampled from a normal distribution with mean μ_i and standard deviation $\frac{\sigma_i}{\sqrt{n_i}}$ (uncertainty on the mean). So, the mean is sampled from: $\mu_i \sim \mathcal{N}(\mu_i, \frac{\sigma_i}{\sqrt{n_i}})$ (27)
- σ_i: the sample variance S_i^2 of a sample of size n_i and with σ_i^2 variance is chi-square distributed: $\frac{(n_i-1)S^2}{\sigma_i^2} \sim \chi^2(n_i - 1)$ (27). So, the variance is sampled from: $S_i^2 \sim \frac{\sigma_i^2}{n_i-1}\chi^2(n_i - 1)$

The precise sampling distribution for each input parameters in the triple log-normal case in detailed in Table I (case B).

2.7.3.Comparing risk estimations

As highlighted in section 2.4, other distributions than the log-normal distribution can be used to estimate the risk of allergic reaction. The sample distributions of the parameters' input distributions are then different. The different cases where the uncertainty propagation from the different input distributions to the risk are summarized in Table I. The parameters distributions for each input distribution are also presented in Table I.

Input	Distribution	Case A	Case B	Case C	Case D
Consumption (X)	Parameter	No uncertainty	$\mu_X \sim \mathcal{N}(\mu_X, \frac{\sigma_X}{\sqrt{n_X}})$ $S_X{}^2 \sim \frac{\sigma_X{}^2}{n_X - 1}\chi^2(n_X - 1)$	$\mu_X \sim \mathcal{N}(\mu_X, \frac{\sigma_X}{\sqrt{n_X}})$ $S_X{}^2 \sim \frac{\sigma_X{}^2}{n_X - 1}\chi^2(n_X - 1)$	Bootstrap
	Input	Log-Normal $X \sim \mathcal{LN}(\mu_X, \sigma_X)$	Log-Normal $X \sim \mathcal{LN}(\mu_X, \sigma_X)$	Log-Normal $X \sim \mathcal{LN}(\mu_X, \sigma_X)$	Empirical
Contamination (Y)	Parameter	No uncertainty	$\mu_Y \sim \mathcal{N}(\mu_Y, \frac{\sigma_Y}{\sqrt{n_Y}})$ $S_Y{}^2 \sim \frac{\sigma_Y{}^2}{n_Y - 1}\chi^2(n_Y - 1)$	$\lambda_Y \sim \mathcal{N}(\mu_Y, \frac{\sigma_Y}{\sqrt{n_Y}})$	$\lambda_Y \sim \text{Gamma}(\alpha + n_Y, \beta + \sum_{j=1}^{n_Y} Y_j)$ With vague prior parameters $\alpha=\beta=10^{-3}$
	Input	Log-Normal $Y \sim \mathcal{LN}(\mu_Y, \sigma_Y)$	Log-Normal $Y \sim \mathcal{LN}(\mu_Y, \sigma_Y)$	Exponential $Y \sim \text{Exp}(\lambda_Y)$	Exponential $Y \sim \text{Exp}(\lambda_Y)$
Threshold (Z)	Parameter	No uncertainty	$\mu_Z \sim \mathcal{N}(\mu_Z, \frac{\sigma_Z}{\sqrt{n_Z}})$ $S_Z{}^2 \sim \frac{\sigma_Z{}^2}{n_Z - 1}\chi^2(n_Z - 1)$	$a_Z \sim \mathcal{N}(\widehat{\mu_{a_Z}}, \widehat{\sigma_{a_Z}})$ $b_Z \sim \mathcal{N}(\widehat{\mu_{b_Z}}, \widehat{\sigma_{b_Z}})$	Bayesian inferences with vague priors: $a_Z \sim \text{Gamma}(10^{-3}, 10^{-3})$ $b_Z \sim \text{Gamma}(10^{-3}, 10^{-3})$
	Input	Log-Normal $Z \sim \mathcal{LN}(\mu_Z, \sigma_Z)$	Log-Normal $Z \sim \mathcal{LN}(\mu_Z, \sigma_Z)$	Weibull $Z \sim \text{Weibull}(a_Z, b_Z)$	Weibull $Z \sim \text{Weibull}(a_Z, b_Z)$

TABLE I: SUMMARY OF THE DIFFERENT CASE PRESENTED TO ASSESS THE UNCERTAINTY PROPAGATION FROM THE INPUTS TO THE RISK OF ALLERGIC REACTION

In case A, no uncertainty on the three inputs was integrated in the risk calculation. The calculation is equivalent to the probabilistic risk assessment developed by TNO and described in (7). In case B, the uncertainty on the input parameters was added, so the main sources of uncertainty can be identified. As explained in section 2.5.2, in this case the uncertainty can also be calculated mathematically. Therefore, uncertainty estimated from simulation and mathematical calculation will be compared. In case C, others distributions than the log-normal distribution (case A and B) were used. The parameters are sampled from standard distribution which described the uncertainty on the parameters. The parameter for the exponential distribution is sampled from a normal distribution with similar distribution as the log-normal case. The parameters of the Weibull distribution are estimated by fitting a survival model. Thus, the two

parameters of the Weibull distribution are sampled from a normal distribution with mean parameter: the estimated parameters, and standard deviation parameter: the standard deviation on the parameters estimated in survival regression. In case D, the uncertainties on the parameters are calculated using Bayesian inferences as described in (6).

2.8. Simulations and software

Once the parameters' sampling distribution had been selected, the risk was calculated for each set of parameters. The number of replications (K) to evaluate the uncertainty on the risk was set to 1 000 and the number of iterations (n) per replication was set to 10 000. Based on some investigations some detailed in this paper, these numbers of replications and iterations were found to be the best compromise between the computation time and accuracy. The simulations were performed with the R software version 3.2.2 (28). The threshold distribution was fit to a survival model using the survival package (version 2.38.3) for the frequentist method and the JAGS software version 4.0.0 (29) for the Bayesian method.

3. RESULTS

3.1. Input distributions

3.1.1.Concentration of unattended allergen distribution (Y)

The mean and the standard deviation of the contamination distribution are 77 ppm and 161 respectively. Figure 3 shows the distribution of the contamination distribution with the histogram and the log-normal distribution with mean and variance calculated from the concentration data.

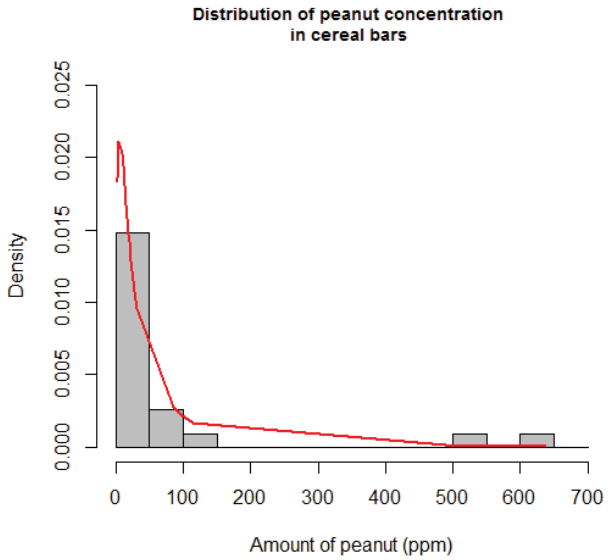


FIGURE 3: DISTRIBUTION OF PEANUT CONCENTRATION IN CEREALS BARS (HISTOGRAM AND FITTED LOG-NORMAL DISTRIBUTION)

Parameter	Mean	SD	P2.5%	Median	P97.5%
μ_Y – case B	3.496	0.273	2.942	3.492	4.042
σ_Y – case B	1.283	0.205	0.904	1.275	1.705
λ_Y – case C	0.013	0.001	0.010	0.013	0.016
λ_Y – case D	0.013	0.003	0.008	0.013	0.019

TABLE II: CONCENTRATION PARAMETERS DISTRIBUTION FOR THE CASES B, C AND D

Summary statistics of the parameters' distribution used to fit the concentration distribution are presented in Table II. μ_Y and σ_Y are the parameters used to fit a log-normal distribution in case B to the concentrations distribution. An exponential distribution is also used in case C and D to describe the concentration distribution: λ_Y is the parameter of the exponential distribution. Table II also shows that the distribution of the λ_Y parameter is similar when Monte Carlo simulations and Bayesian inferences are performed; the average λ_Y is 0.013 in case C and D.

3.1.2.Consumption distribution (X)

With the 350 participants in the three combined Food Consumption surveys consuming cereal bars, the mean and the standard deviation of the largest portion of cereal bar consumed per consumer were

calculated and equal to 32g and 28g respectively. Furthermore, 95% of the consumption lies between 1.9 g and 291.0g. Figure 4 shows the distribution of the consumption distribution with the histogram and the log-normal distribution with mean and variance calculated from the consumption data.

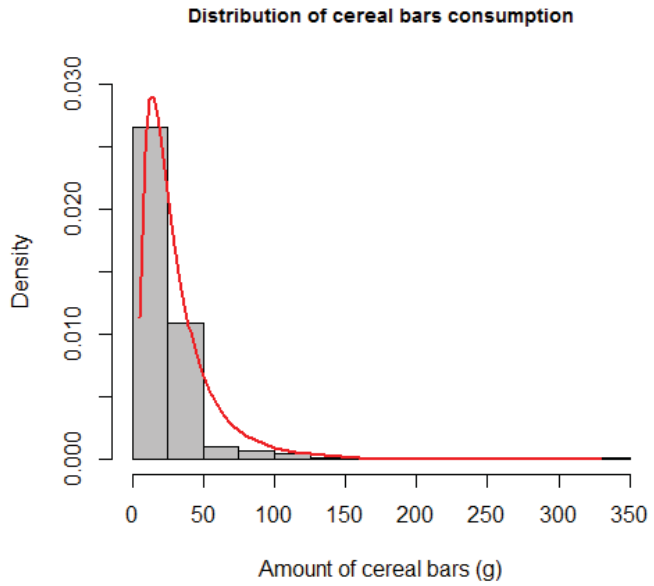


FIGURE 4: DISTRIBUTION OF CEREAL BARS CONSUMPTION (HISTOGRAM AND FITTED LOG-NORMAL DISTRIBUTION)

A log-normal distribution is used in cases B and C to describe the consumption distribution. Summary statistics of its parameters' (μ_X and σ_X) distribution are presented in Table III.

Parameter	Mean	SD	P2.5%	Median	P97.5%
μ_X – case B	-3.719	0.039	-3.796	-3.719	-3.642
σ_X – case B	0.747	0.028	0.691	0.746	0.801

TABLE III: CONSUMPTION PARAMETERS DISTRIBUTION FOR THE CASES B AND C

3.1.3.Challenge threshold distribution (Z)

The summary statistics of the parameters of the Log-Normal and Weibull distribution parameters fitted with a survival model are presented in Table IV.

Parameter	Mean	SD	P2.5%	Median	P97.5%
μ_z – case B	4.088	0.238	3.612	4.088	4.55
σ_z – case B	2.987	0.171	2.668	2.984	3.324
a_z – case C	0.382	0.027	0.331	0.381	0.438
b_z – case C	229.621	55.099	142.773	223.298	351.067
a_z – case D	0.379	0.027	0.328	0.379	0.433
b_z – case D	225.075	51.906	139.852	219.042	341.439

TABLE IV: THRESHOLD PARAMETERS DISTRIBUTION FOR THE CASES B, C AND D

The distribution of μ_z and σ_z, the parameters of the log-normal distribution used to fit the threshold distribution (case B), is summarized with the mean and standard deviation, median and the 2.5% and the 97.5% quantiles. The averages estimated coefficient for μ_z is 4.09 and for σ_z is 2.98. The distribution of a_z and b_z, the parameters of the Weibull distribution used to fit the threshold distribution, was calculated in the frequentist (case C) and Bayesian (case D) way. In Table IV, it can be seen that the distribution of the parameters a_z and b_z are similar. The mean distribution of a_z is 0.38 in case c and 0.38 in case D. And the mean distribution of b_z is 229.6 in case C and 225.1 in case D.

The cumulative probability of reaction estimated with the survival package for the log-normal and the Weibull distribution are plotted on Figure 5, the 95% confidence interval is also indicated for each distribution.

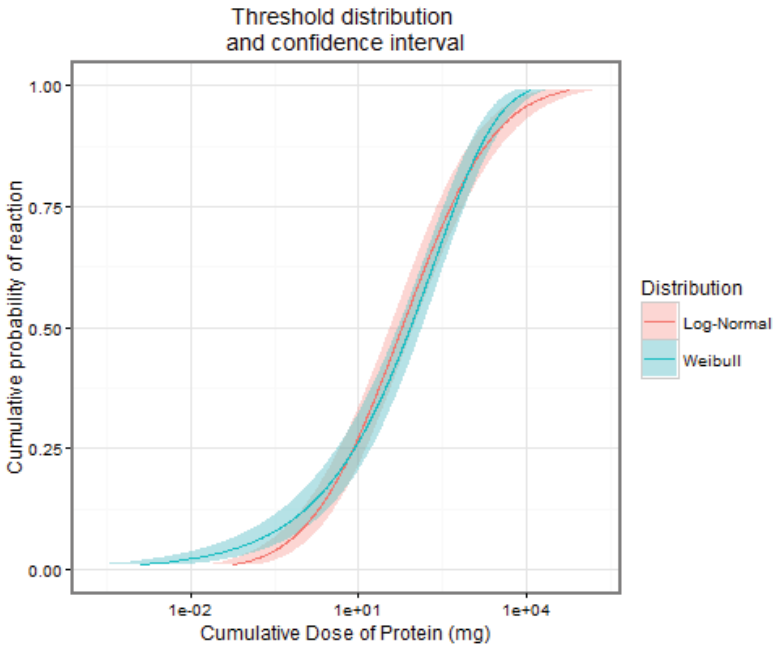


FIGURE 5: FITTED THRESHOLD DISTRIBUTION WITH LOG-NORMAL AND WEIBULL DISTRIBUTION AND ITS 95% CONFIDENCE INTERVAL

3.2. Model comparison

Case	Mean	SD	2.5%	Median	97.5%
Case A	9.79%	Not relevant	Not relevant	Not relevant	Not relevant
Case B - calculation	9.91%	2.30%	5.70%	9.79%	14.41%
Case B - simulation	9.93%	2.32%	5.82%	9.80%	14.52%
Case C	13.76%	2.09%	9.94%	13.65%	17.97%
Case D	14.69%	2.31%	10.48%	14.66%	19.41%

TABLE V: RISK DISTRIBUTION ESTIMATION FOR THE 4 CASES

In the peanut allergic population consuming cereal bars, the probability of allergic reaction after consuming a contaminated cereal bar range from 9.79% to 14.69%, depending on the way the risk of allergic reaction is estimated. When three log-normal distributions are used (case A and B), the average risk of allergic reaction is similar: around 9.8-9.9%. In case B, the risk's distribution can be calculated either with mathematical calculation or with simulations. In those cases, the risks distributions are very alike as it can be noticed in Table V. The average probability of allergic reaction was actually calculated as 9.91% (CI95: 5.70%-14.41%) and simulated as 9.95% (CI95: 5.82%-14.52%), which considering the

simulation error is identical. In case C and D, other distributions, than log-normal distributions, were used to estimate the risk of allergic reaction. That's why the average risk of allergic reaction is different (higher in this case) than in case A and B. However, the risks estimated are similar when using the same distributions for the input parameters but different simulations methods (i.e. second order Monte Carlo simulations in case C and Bayesian inferences in case D). The average risk calculated are both around 14%: in case C, the risk in on average 13.76% (CI95: 9.94%-17.97%) and in case D, the risk in on average 14.69% (CI95: 10.48%-19.41%). Thus, Table V shows that also in case C and D the risks distribution are similar and that no differences between simulations method can be highlighted.

3.3. Uncertainty propagation

Uncertainty on parameters	Case B		Case C	Case D
	Calculation	Simulation		
Consumption (X)	$0.43\ 10^{-03}$	$1.47\ 10^{-03}$	$1.68\ 10^{-03}$	$1.12\ 10^{-03}$
Contamination (Y)	$2.23\ 10^{-02}$	$2.29\ 10^{-02}$	$0.35\ 10^{-02}$	$1.14\ 10^{-02}$
Threshold (Z)	$2.85\ 10^{-02}$	$2.95\ 10^{-02}$	$4.04\ 10^{-02}$	$4.07\ 10^{-02}$
Sum of uncertainty from individual parameters	$5.13\ 10^{-02}$	$5.39\ 10^{-02}$	$4.57\ 10^{-02}$	$5.32\ 10^{-02}$
All parameters	$5.28\ 10^{-02}$	$5.37\ 10^{-02}$	$4.36\ 10^{-02}$	$5.37\ 10^{-02}$

TABLE VI: RISK'S STANDARD DEVIATION (IN %) WHEN UNCERTAINTY IS ADDED TO INPUTS' PARAMETERS INDIVIDUALLY AND ALL AT THE SAME TIME

Table VI presents the risks standard deviation for cases presented in Table I and when the uncertainty from the input parameters are added individually and then all together. In all four cases, the uncertainty on the risk from the three input distributions included individually add-up to the uncertainty when the uncertainty from the three input distributions is added simultaneously. Thus, it can be verified that the assumption that the three input distribution contributes independently to the uncertainty on the risk is correct. Moreover, the contributions of each individual source of uncertainty are ranked in the order for all the cases presented in Table VI, regarding the data used in this example. In case C and D, the principal

source of uncertainty is the threshold with a standard deviation around 0.04%. In case B, the uncertainty of thresholds is comparable to the uncertainty of the exposure (*consumption X contamination*). It is mainly due to the difference in threshold modeling: in case B, the log-normal distribution is used, and in case C and D, the Weibull distribution is used. Consumption and contamination distribution contribute to the same level to the uncertainty on the risk. The contamination distribution has a standard deviation ranging from 0.003% to 0.02%, depending on the case. And finally consumption distribution contributes to the uncertainty on the risk with a standard deviation ranging from 0.004% to 0.02%.

4. CONCLUSION/DISCUSSION

In this paper, the two methods widely used to estimate the risk of allergic reaction had been compared. It had been highlighted that the method in (7) do not actually integrate the uncertainty in the risk estimation. It is then recommended to not communicate the uncertainty on the risk when estimating the risk of allergic reaction with this model, as it only reflects the simulation error. Then, the comparison between the Bayesian and non-Bayesian way of estimating the risk of allergic reaction has not proven any difference in the risk estimation and the risk uncertainty. However, an easier mathematical expression and software implementation of the frequentist model suggest a preferable use of this model.

In the triple-log-normal case, the risk can be easily expressed mathematically with a normal distribution with the three input distribution parameters. That's why it was decided not to include other parameters, as the chance of contamination or the prevalence of peanut allergy, in the risk estimation. However, the models presented in this paper can be extended by integrating other variable in the risk calculation. The extension was not presented in this paper for a purpose of clarity. Furthermore, additional factors like the severity of the allergic reaction can be added to the risk estimation. Thus, the threshold distribution

could be distinguished into mild and severe (7) and the severity of the reaction could also be predicted if there is sufficient data in the two groups.

It can be concluded from the uncertainty analysis that improvement on data should be done as a priority on the threshold data. When we compare the uncertainty coming from the exposure inputs distributions (*consumption X contamination*) with the one coming from the threshold input distribution, the order of magnitude is comparable. It is actually expected due to the way the risk is calculated: the threshold is compared to multiplication of the consumption by the contamination level to calculate the risk. For example, the conclusion made from the uncertainty analysis might differ if we use data different data for threshold of concentration distributions. Only peanut threshold data publically available were used in the risk assessment. However, more data are available within organizations specialized in allergen risk assessment (10). Thus, the main source of uncertainty might be different than the threshold distribution. This will actually help to identify for which distribution the quality of the distribution should be improved in order to have better accuracy of the risk estimation. On the other hand a data set from 158 peanut challenges was used. This is a high number of challenges compared to some of the other foods where a challenge distribution has been described (10). Furthermore, the data quality available is different for the different allergens (30). Thus, similar uncertainty analysis could be run for all the allergens for which probabilistic risk assessment is usually run. And, the allergen for which the data quality should be improved as a priority could be identified.

Variability (i.e. heterogeneity among individuals) was not investigated in this paper. Besides that further calculation was integrated in the method developed in (6) (equivalent to case D). Thus, as shown in (6), heterogeneity of the risk of allergic reaction can be investigated at the consumers' level. However, for each individual consuming cereal bars the data collection scheme does not allow the acquisition of the matching threshold of reaction. The consumption and threshold data are actually collected in different

studies. That's why, the individuals variabilities were not found relevant to be calculated in this paper. However, the recent MIRABEL study (31), registered for allergic individuals their consumption of certain food products and their corresponding threshold of reaction. With such data the variability at individuals level could be calculated using a complete Bayesian network (32).

Sensitivity analysis had already been carried out with the model developed in (7) (33). With sensitivity analysis, the input variables that have the most influence on the model output (i.e. the risk of allergic reaction) were identified (24). When the threshold distribution is shifted for simulating a more potent allergen, the estimation of the risk of allergic reaction increased a lot. It was highlighted in this article that model definition could not allow correct evaluation of the uncertainty on the risk. However, it is expected that the rank of the influence of the input variables on the risk to be the same.

In conclusion we have compared the two published methods in food allergy probabilistic risk assessment and found that although different mathematical formulation have been used overall the results obtained are very similar. In addition we have proposed a frequentist method that is able to propose a full uncertainty approach, so the additional data will most efficiently contribute to reduce uncertainty can be identified for the different allergen.

Appendix 1: partial derivative calculation within the uncertainty propagation formula

Each derivative can be then calculated separately:

$$\frac{\partial p_u}{\partial \mu_x^L} = \phi\left(\frac{\mu_z^L - \mu_x^L - \mu_y^L}{\sqrt{(\sigma_x^L)^2+(\sigma_y^L)^2+(\sigma_z^L)^2}}\right) \cdot \frac{1}{\sqrt{(\sigma_x^L)^2+(\sigma_y^L)^2+(\sigma_z^L)^2}}$$

$$\frac{\partial p_u}{\partial \mu_y^L} = \phi\left(\frac{\mu_z^L - \mu_x^L - \mu_y^L}{\sqrt{(\sigma_x^L)^2+(\sigma_y^L)^2+(\sigma_z^L)^2}}\right) \cdot \frac{1}{\sqrt{(\sigma_x^L)^2+(\sigma_y^L)^2+(\sigma_z^L)^2}}$$

$$\frac{\partial p_u}{\partial \mu_z^L} = -\phi\left(\frac{\mu_z^L - \mu_x^L - \mu_y^L}{\sqrt{(\sigma_x^L)^2+(\sigma_y^L)^2+(\sigma_z^L)^2}}\right) \cdot \frac{1}{\sqrt{(\sigma_x^L)^2+(\sigma_y^L)^2+(\sigma_z^L)^2}}$$

$$\frac{\partial p_u}{\partial (\sigma_x^L)^2} = \phi\left(\frac{\mu_z^L - \mu_x^L - \mu_y^L}{\sqrt{(\sigma_x^L)^2+(\sigma_y^L)^2+(\sigma_z^L)^2}}\right) \cdot \frac{\mu_z^L - \mu_x^L - \mu_y^L}{2\left(\sqrt{(\sigma_x^L)^2+(\sigma_y^L)^2+(\sigma_z^L)^2}\right)^3}$$

$$\frac{\partial p_u}{\partial (\sigma_x^L)^2} = \frac{\partial p_u}{\partial (\sigma_y^L)^2} = \frac{\partial p_u}{\partial (\sigma_z^L)^2}$$

BIBILIOGRAPHIE

1. Gendel SM. Comparison of international food allergen labeling regulations. Regul Toxicol Pharmacol. Elsevier Inc.; 2012;63(2):279–85.

2. Fernández-Rivas M, Asero R. Risk Management for Food Allergy. Madsen CB, W. R. Crevel R, Mills C, L. Taylor S, editors. Risk Management for Food Allergy. Elsevier; 2014. 25-43 p.

3. DunnGalvin A, Chan CH, Crevel R, Grimshaw K, Poms R, Schnadt S, et al. Precautionary allergen labelling: Perspectives from key stakeholder groups. Allergy Eur J Allergy Clin Immunol. 2015;70(9):1039–51.

4. Barnett J, Muncer K, Leftwich J, Shepherd R, Raats MM, Gowland MH, et al. Using "may contain" labelling to inform food choice: a qualitative study of nut allergic consumers. BMC Public Health. BioMed Central Ltd; 2011;11(1):734.

5. Afssa. Food allergies and advisory labelling. 2008;(November).

6. Rimbaud L, Heraud F, La Vieille S, Leblanc JC, Crepet A. Quantitative risk assessment relating to adventitious presence of allergens in food: A probabilistic model applied to peanut in chocolate. Risk Anal. 2010;30(1):7–19.

7. Spanjersberg MQI, Kruizinga a. G, Rennen M a J, Houben GF. Risk assessment and food allergy: the probabilistic model applied to allergens. Food Chem Toxicol. 2007;45(1):49–54.

8. Remington BC, Baumert JL, Marx DB, Taylor SL. Quantitative risk assessment of foods containing peanut advisory labeling. Food Chem Toxicol. Elsevier Ltd; 2013;62:179–87.

9. Birot S, Madsen CB, Kruizinga AG, Crépet A, Christensen T, Brockhoff PB. Combining food consumption data from different countries for creating food groups for allergen risk assessment

(in Europe). Manuscript submitted for publication. 2016;

10. Taylor SL, Baumert JL, Kruizinga AG, Remington BC, Crevel RWR, Brooke-Taylor S, et al. Establishment of Reference Doses for residues of allergenic foods: Report of the VITAL Expert Panel. Food Chem Toxicol. Elsevier Ltd; 2014;63:9–17.

11. Taylor SL, Crevel RWR, Sheffield D, Kabourek J, Baumert J. Threshold dose for peanut: Risk characterization based upon published results from challenges of peanut-allergic individuals. Food Chem Toxicol. Elsevier Ltd; 2009;47(6):1198–204.

12. Anagnostou K, Islam S, King Y, Deighton J, Clark AT, Ewan PW. British Society for Allergy and Clinical Immunology Annual Conference 2009 Abstracts. British Society for Allergy and Clinical Immunology Annual Conference 2009 Abstracts. Blackwell Publishing Ltd; 2009. p. 1937–58.

13. Atkins FM, Steinberg SS, Metcalfe DD. Evaluation of immediate adverse reactions to foods in adult patients: II. A detailed analysis of reaction patterns during oral food challenge. J Allergy Clin Immunol. 1985;75(3):356–63.

14. Blumchen K, Ulbricht H, Staden U, Dobberstein K, Beschorner J, de Oliveira LCL, et al. Oral peanut immunotherapy in children with peanut anaphylaxis. J Allergy Clin Immunol. 2010;126(1):83–91.e1.

15. Clark AT, Ewan PW. Good prognosis, clinical features, and circumstances of peanut and tree nut reactions in children treated by a specialist allergy center. J Allergy Clin Immunol. 2008;122(2):286–9.

16. Hourihane JO, Kilburn SA, Nordlee JA, Hefle SL, Taylor SL, Warner JO. An evaluation of the sensitivity of subjects with peanut allergy to very low doses of peanut protein: A randomized, double-blind, placebo-controlled food challenge study. J Allergy Clin Immunol. 1997;100(5):596–

600.

17. Leung DYM, Sampson HA, Yunginger JW, Burks AW, Schneider LC, Wortel CH, et al. Effect of Anti-IgE Therapy in Patients with Peanut Allergy. N Engl J Med. Massachusetts Medical Society; 2003 Mar 13;348(11):986–93.

18. Lewis SA, Grimshaw KEC, Warner JO, Hourihane JO. The promiscuity of immunoglobulin E binding to peanut allergens, as determined by Western blotting, correlates with the severity of clinical symptoms. Clin Exp Allergy. Blackwell Publishing; 2005;35(6):767–73.

19. NELSON H, LAHR J, RULE R, BOCK A, LEUNG D. Treatment of anaphylactic sensitivity to peanuts by immunotherapy with injections of aqueous peanut extract1. J Allergy Clin Immunol. Mosby; 1997 Jun;99(6):744–51.

20. Nicolaou N, Poorafshar M, Murray C, Simpson A, Winell H, Kerry G, et al. Allergy or tolerance in children sensitized to peanut: Prevalence and differentiation using component-resolved diagnostics. J Allergy Clin Immunol. 2010;125(1):191–197.e13.

21. Oppenheimer JJ, Nelson HS, Bock SA, Christensen F, Leung DYM. Treatment of peanut allergy with rush immunotherapy. J Allergy Clin Immunol. 1992;90(2):256–62.

22. Patriarca G, Nucera E, Pollastrini E, De Pasquale T, Lombardo C, Buonomo A, et al. Oral Rush Desensitization in Peanut Allergy: A Case Report. Dig Dis Sci. 2006;51(3):471–3.

23. Wainstein BK, Studdert J, Ziegler M, Ziegler JB. Prediction of anaphylaxis during peanut food challenge: usefulness of the peanut skin prick test (SPT) and specific IgE level. Pediatr Allergy Immunol. Blackwell Publishing Ltd; 2010;21(4p1):603–11.

24. Makowski D. Uncertainty and sensitivity analysis in quantitative pest risk assessments; practical

rules for risk assessors. NeoBiota. 2013;18:157–71.

25. EFSA Scientific Committee. Guidance on Uncertainty in EFSA Scientific Assessment. The EFSA Journal.

26. Navidi W. Statistics for engineers and scientists. Boston, Mass: McGraw-Hill; 2006.

27. Johnson R. Miller & Freund's probability and statistics for engineers. Boston: Prentice Hall; 2011.

28. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2015.

29. Plummer M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. Proceedings of the 3rd International Workshop on Distributed Statistical Computing. 2003.

30. Madsen CB, Houben G, Hattersley S, Crevel RWR, Remington BC, Baumert JL. Risk Management for Food Allergy. Risk Management for Food Allergy. Elsevier; 2014. 101-126 p.

31. Crépet A, Papadopoulos A, Elegbede CF, Ait-Dahmane S, Loynet C, Millet G, et al. Mirabel: An integrated project for risk and cost/benefit analysis of peanut allergy. Regul Toxicol Pharmacol. 2015;71(2):178–83.

32. Albert I, Grenier E, Denis JB, Rousseau J. Quantitative risk assessment from farm to fork and beyond: A global Bayesian approach concerning food-borne diseases. Risk Anal. 2008;28(2):557–71.

33. Kruizinga AG, Briggs D, Crevel RWR, Knulst AC, Bosch LMC van den, Houben GF. Probabilistic risk assessment model for allergens in food: sensitivity analysis of the minimum eliciting dose and food consumption. Food Chem Toxicol. 2008;46(5):1437–43.

Chapter 5

Shiny application: estimation of risk of allergic reaction

One aim of the project was to design a more user friendly interface for assessing the risk of allergic reaction with probabilistic modelling. So, the risk assessment could be performed easily by non-statistician. It was then decided to develop a Shiny application, which is an interactive web application framework for R. R has the advantage to be an open-source software and dynamically developed that can be used for statistical computing and graphics. And, Shiny facilitates the use of the R software through the interactive interface. Thus, the risk assessment application can be used by the iFAAM Work Package partners without statistical and software training.

In this chapter, the input dataset to the risk assessment R function will be first presented along with how the way they are handle within the Shiny application. Then, the methodological choices for the risk computation will be described regarding the investigation made in chapter 4. Finally, the output results will be explained. The R code for the Shiny application with the risk function used to estimate the risk of allergic reaction are can be found appendixC.

5.1 Risk calculation

Several inputs are required for the allergen risk assessment. Depending for which population the risk of allergic reaction is calculated, different input are integrated in the risk calculation. As explained in chapter 1, three different kind a inputs regarding the consumption of the food products, the contamination of the food products and the challenge data are used to calculate the risk of allergic reaction in the allergic user of a specific food product population, in the allergic population or in the overall population.

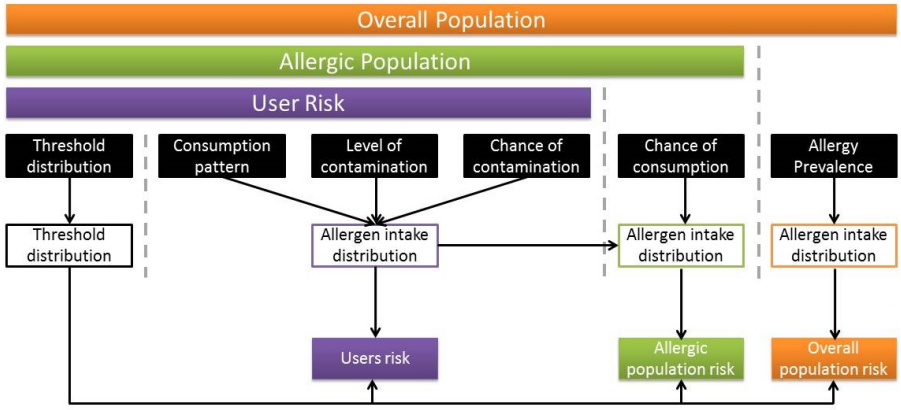


Figure 5.1: Inputs' summary for the risk of allergic reaction calculation for the different populations

On figure 5.1, the input that are used to calculate the risk of allergic reaction for the different populations are detailed. Thus, the risk of allergic reaction in **the allergic products users population** requires four inputs. The threshold distribution, the consumption pattern of the specific food product and regarding the contamination, two inputs are needed: the level of concentration in the food product (i.e allergen concentration) and the chance that the food product is contaminated. Furthermore, the risk of allergic reaction in **the allergic population** is calculated adding the information the chance of consuming the product in the population. Finally, the risk of allergic reaction is calculated in **the overall population** by adding the prevalence of allergy in the population (i.e the percentage of the population suffering from allergy with the selected allergen). The three risks are calculated within the shiny with the standard data in the Shiny application and the data provided by the user. The data and their relation to the risk estimation are described in the following sections.

5.2 Data on the consumption

The consumption is described with two inputs. The chance of consumption, i.e. what is the proportion of persons consuming a food product. And the level of consumption, i.e how much is the consumption when the food product is actually eaten.

Consumption surveys from Netherlands, France and Denmark were shared by the iFAAM work package partners. As explained in chapter 3, the consumers' consumption was recorded over 7 or 2 days, depending on the country. The maximum consumption on an eating occasion was calculated for each consumer and each food item. For each country, consumption amounts were stored in a table $consumer \times fooditem$. The tables were then merged and used as standard dataset within the Shiny application. The country information is also indicated for each consumer.

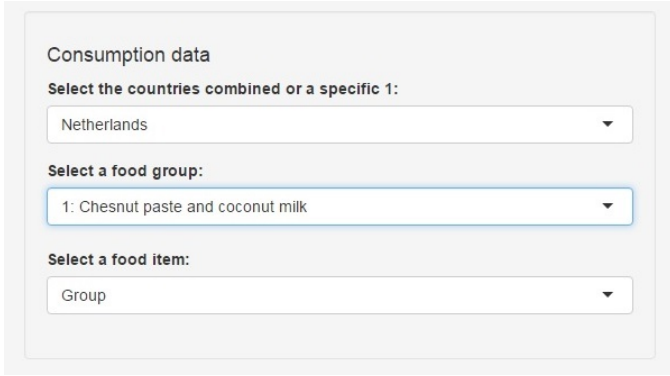


Figure 5.2: User's choice for the consumption inputs

Figure 5.2 shows the options that the user can choose regarding the consumption data. First, the country for which the risk assessment is performed is selected among Netherlands, France, Denmark or the three countries combined. And the food group and food item in the selected country for which the risk assessment is performed can also be selected.

5.2.1 Food groups for allergen risk assessment

In chapter 3, 56 food groups for allergen risk assessment were designed. Then, it make sense to re-use those groups for the Shiny application. The food items used in the three surveys were assigned to a food group according to the level of consumption and expert knowledge. The information on food items' corresponding food group is stored in file. When a country and food group is selected, the corresponding food items are displayed. So the user can choose to perform the risk assessment either for the group of food items or for a specific food item within the selected food group. Information about the food group name and the food items names are also displayed, so the user is informed of which food items can be selected for performing the risk assessment. Then the consumption dataset included in the Shiny application is filtered accordingly to the food group and the country selected. Thus, data on the consumption of the food products for the selected are used to calculate the chance of consumption and the level of consumption inputs for the risk function.

5.2.2 Chance of consumption

Once the food group or the food item is selected, the chance of consumption mean and standard deviation are calculated from consumption survey with bootstrap. The food group or food item consumers are re-sampled 1000 times according to surveys sampling weights. Then, the mean and standard deviation of the consumption chance are calculated from the distribution of consumption calculated with consumers bootstrap.

5.2.3 Level of consumption

The level of consumption's distribution is directly used by the risk function. If a food item is selected the non-zero consumptions are filtered out. Thus, this distribution with corresponding sampling weights are used. Or if a food groups is selected, the distribution of the food items consumed are stacked and used jointly, still with corresponding consumers' sampling weights and still with the non-zero consumptions.

5.3 Data on the contamination

The data on the contamination needs to be filled out by the user, as the risk assessment is performed case by case.

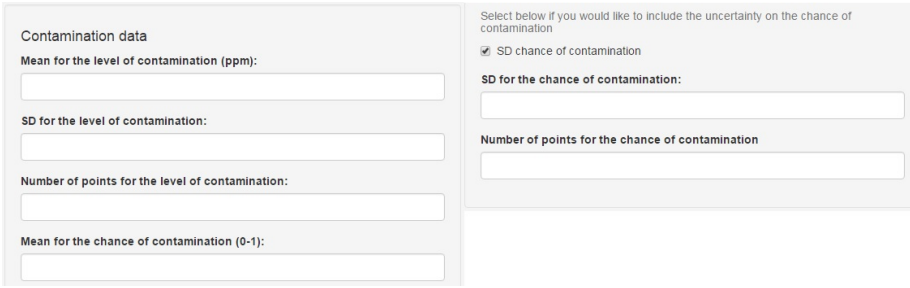


Figure 5.3: User's choice for the contamination inputs

In order to integrate the uncertainty from the contamination measurement to the risk of allergic reaction, different figures are filled out by the user. On figure 5.3, the information on the level and chance of contamination needed for the estimation of the risk of allergic reaction is detailed.

5.3.1 Level of contamination

The mean concentration of allergen in the contaminated food product needs to be fill out in ppm. Furthermore, other information on the number of measurement made to estimate the mean concentration of allergen and the standard deviation of the concentration of allergen are given by the user.

5.3.2 Chance of contamination

The mean chance of contamination is also supplied by the user. However, there is an option for including or not the parameter uncertainty in the risk uncertainty. When only one set of contaminated products was analysed, repeated concentration measurement for estimating the chance of contamination is not always possible. So, the uncertainty coming from this parameter cannot be included in the risk calculation. However, if several measurement of the chance of contamination can be made, it is still possible to include the uncertainty from

the parameter by checking a box that enables the user to fill the number of points and the standard deviation of the chance of contamination.

5.4 Data on the threshold and the prevalence of food allergy

The last information is the data regarding the threshold and the prevalence of food allergy.

Figure 5.4: User's choice for the threshold inputs

On figure 5.4, the different allergens and populations that can be selected by the user are displayed. The threshold database actually records information on for which allergen and population the threshold are measured in the clinical studies. The corresponding allergens and populations prevalences in the population are stored in the prevalence dataset for each country.

5.4.1 Threshold data

The threshold values (NOAEL and LOAEL) for each individuals and each allergen are stored in one dataset. The user selects the allergen for which the risk assessment is performed. And the subset of the threshold dataset is used as an input to the risk assessment function. The threshold database can be updated with NOAEL and LOAEL from new patients and for new allergen within the same structure. As it is an interactive application, the allergen list is automatically updated and the new allergen can be selected by the user to perform a probabilistic risk assessment.

5.4.2 Prevalence of food allergy

The mean, standard deviation and the number of points for estimating the prevalence of allergic reaction for different allergen in different country were collected from scientific publications [12]. The prevalence corresponding to the allergen, population and country chosen by the user is selected in the prevalence dataset and are used as an input to the risk assessment function. The prevalence dataset can be updated when new information on the prevalence in different country and for different allergen become available.

5.5 Methodological choice for risk computation

Once the country, the allergen and the food product or group is selected and the data shaped for the risk function, the risk of allergic reaction can be calculated. The risk simulations are performed based on the methodological investigation made in chapter 4. Thus, second order Monte-Carlo simulations will used to estimate the risk of allergic reaction for the different populations. The standard distributions used to estimate the distribution parameters are defined in this section.

5.5.1 Parameters sampling distributions

As explained in chapter 4, the mean (μ), standard deviation (σ) and coefficient estimates (i.e. a) are sampled from well-known distributions [8]:

- μ: the mean is sampled form a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ (uncertainty on the mean). This parameter's sampling distribution is used either for the mean parameter of a log-normal distribution or for describing the probability parameter's distribution for a binomial distribution.
- σ: the sample variance S^2 of a sample of size n and with σ^2 variance is chi-square distributed: $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$. So the variance is sampled from: $S^2 \sim \frac{\sigma^2}{n-1}\chi^2(n-1)$. This distribution is used to describe the standard deviation parameter of a log-normal distribution.
- $\hat{a}$: the distribution fitted to the threshold data with a survival model use three different distribution: log-normal, log-logistic and Weibull. The

distributions' coefficients are mean and standard deviation are estimated with the **Survival** package in the R software. Thus, the coefficient are described with a normal distribution with mean $\widehat{\mu_a}$ and standard deviation $\widehat{\sigma_a}$.

5.5.2 Summary of the fitted distributions to the inputs

The distribution for all the risk assessment inputs distributions and its parameters distributions are summarized on figure 5.5.

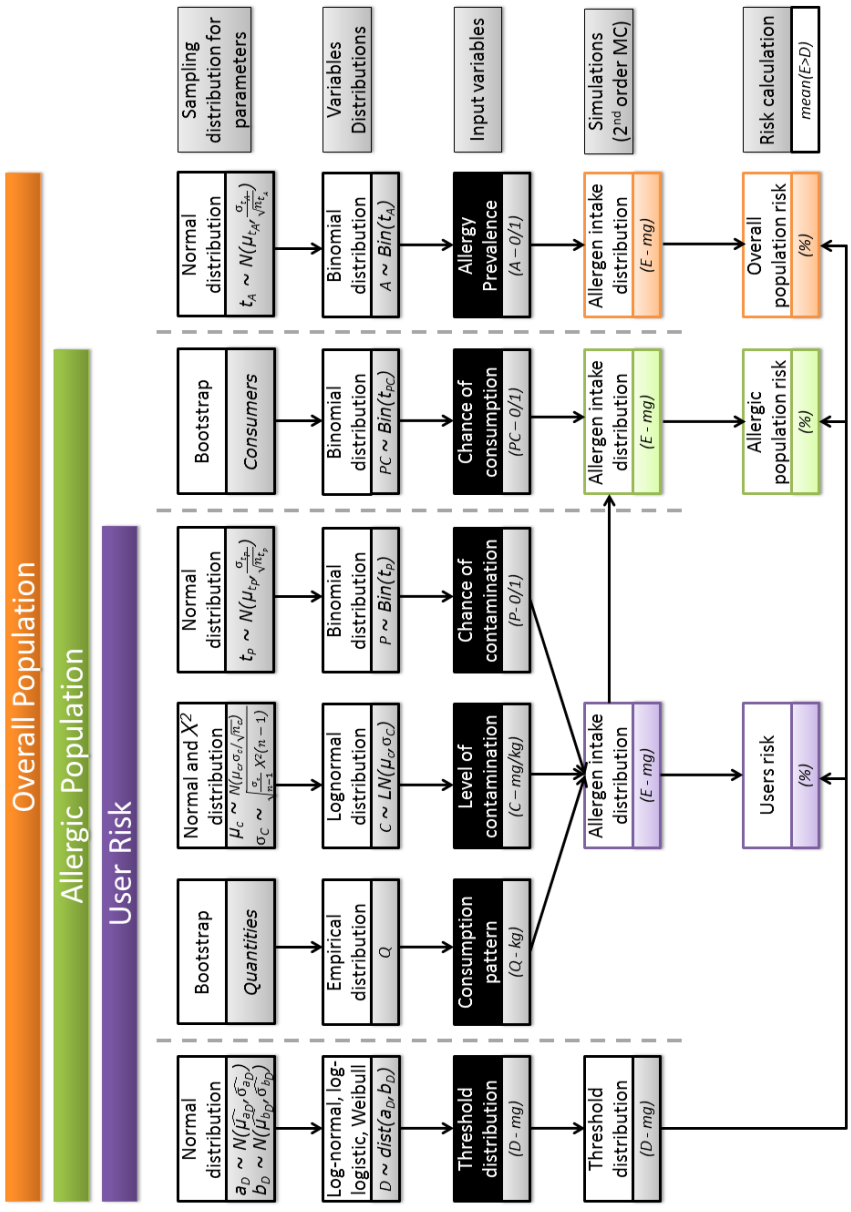


Figure 5.5: Distribution used for the inputs and its parameters

5.5.3 Number of replications and iterations

$N = 10000$ simulations are performed with each set of parameters sampled in their distribution as described in section 5.5.2. In order to evaluate the risk's uncertainty , $K = 1000$ replications are performed with different set of parameters. Investigations on how many replications are needed to balance computation time and accuracy were made and reported in appendix YY. Thus, the simulation error becomes smaller and the risk's confidence interval represents mainly the uncertainty on the risk.

5.6 Results description

The outcome of the risk assessment is summarized with tables and a graph, which will be described in this section.

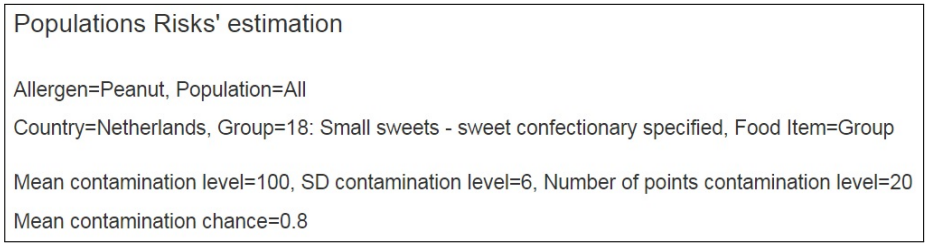
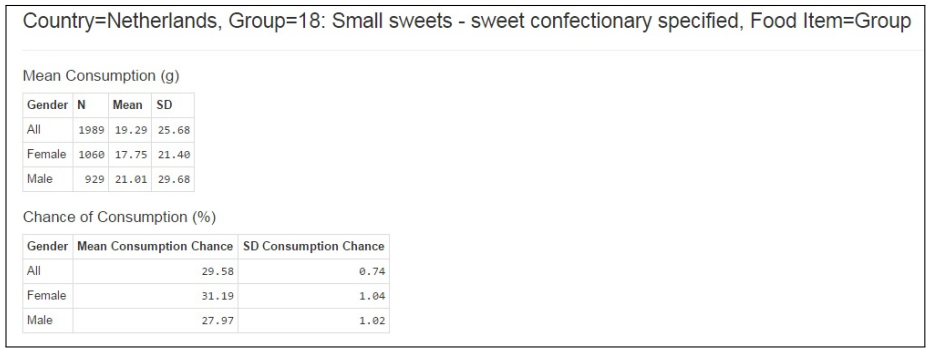


Figure 5.6: Risk assessment allergen, population, food group and contamination information

The food group, the allergen and the population for which the risk assessment is performed are reminded. On figure 5.6, the risk assessment is performed for peanut in the population including both children and adults for the "small sweets - sweet confectionery specified" in Netherlands. Furthermore, the numbers related to the level and chance of contamination filled out by the user are also displayed, so the user can check that no mistakes were made when filling out these numbers. For this example, the risk assessment is run for a mean contamination level of 100 ppm, with 6 ppm of standard deviation calculated with 20 points. And the mean chance of contamination is 80% with no uncertainty integrated from this parameter as the standard deviation and the number of data points were not filled out by the user.

5.6.1 Information on chance and level of consumption

In order to help the user in the risk assessment evaluation, some summary statistics about the level and chance of contamination are indicated.


Country=Netherlands, Group=18: Small sweets - sweet confectionary specified, Food Item=Group

Mean Consumption (g)

Gender	N	Mean	SD
All	1989	19.29	25.68
Female	1060	17.75	21.40
Male	929	21.01	29.68

Chance of Consumption (%)

Gender	Mean Consumption Chance	SD Consumption Chance
All	29.58	0.74
Female	31.19	1.04
Male	27.97	1.02

Figure 5.7: Level and chance of consumption summary statistics for the selected group

Thus, the "small sweets - sweet confectionery specified" are consumed by 1989 in Netherlands, with a mean of 19g ans standard deviation of 26g. This information is also indicated for male and female separated. Furthermore, "small sweets - sweet confectionery specified" are consumed by around 30% of the Dutch population with a standard deviation of around 0.7%. The chance of consumption is also calculated for male and female separately.

5.6.2 Risk of allergic reaction tables

The risk of allergic reaction is calculated for different populations and given with confidence interval.

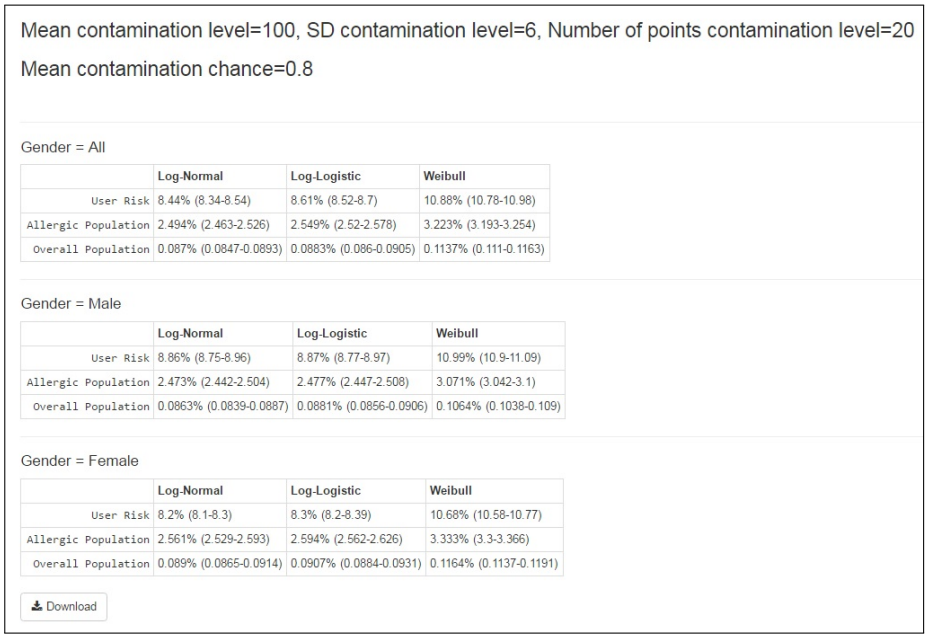


Figure 5.8: Risks table calculated for the input selected by the user

The risk is estimated for three population with three different distribution used to fit the threshold distribution (figure 5.8). Thus, 9 different risks are displayed in the tables for the different combinations of population and distribution. The risk calculated for the allergic risk user population with the log-normal distribution for the example illustrating the Shiny application outputs is 8.40% with a 95% confidence interval: 8.34%-8.54%. The risks of allergic reaction are also calculated for male and female separately. The same table is then displayed for both gender. It also is possible to download the three summary tables with the risk and their confidence interval by clicking the "Download" button.

5.6.3 Predicted reaction graph

In order to assess the number of the allergic reactions and the consequences of the allergen contamination, the individual threshold simulated is plotted against the consumption amount simulated when an allergic reaction happens. Figure 5.9 illustrates the graphical outcome of the risk assessment with the "small sweets - sweet confectionery specified" example and the level of contamination indicated in the previous sections. Summary statistic of the consumption and threshold

distributions are also displayed on figure 5.9. Thus, the average consumption and the 90th percentile of the "small sweets - sweet confectionery specified" consumption distribution is calculated and plotted as horizontal lines. Furthermore, the lowest LOAEL , the ED05 and the ED10 are also calculated from the threshold database, and plotted as vertical lines. The ED05 is the allergen dose to which 5% of the allergic persons react, and ED10 is the allergen dose to which 10% of the allergic persons react.

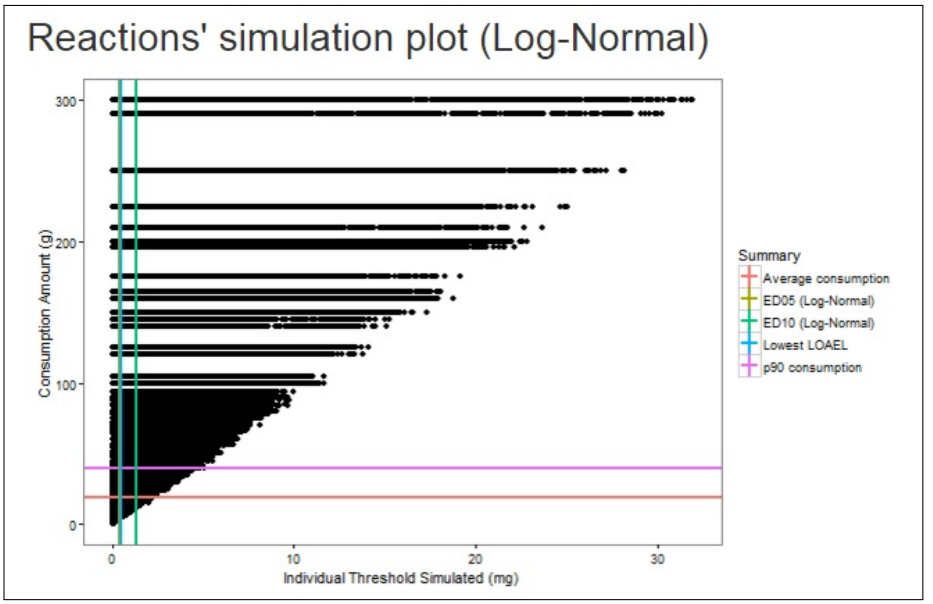


Figure 5.9: Simulated reaction: relation between consumption amounts and individual thresholds simulated

5.6.4 Number of allergic reaction in the different populations

The number of allergic reaction for a specific product can be estimated based on the risk of allergic reaction calculated for the selected food group or food items and the contamination information filled out by the user. After the risk calculation for the selected food group or food items, the product market within the food group and the number of products implicated are filled out by the user. Then, the number of allergic reaction predicted for the different population are calculated by multiplying the risk for the different population and distribution with the market share and number of products. The confidence interval is not

indicated as no uncertainty on the market share and the number of product is filled out by the user. Hence, more investigations on how to obtain the uncertainty on these inputs is needed, so this information can be directly integrated in the risk estimation function.

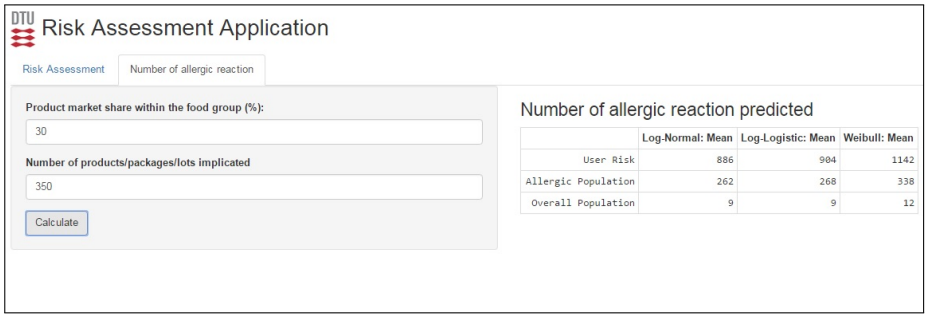


Figure 5.10: Number of allergic reaction predicted for the different population

On figure 5.10, the number of allergic reaction in the different populations are calculated with illustrative example in the "small sweets - sweet confectionery specified" food groups used to illustrate how the risk assessment application is working. The market share of the product within the food group is set at 30% and 350 products are implicated in the incident. Using the log-normal distribution, 886 reaction are predicted for the allergic population user of "small sweets - sweet confectionery specified", 262 for the allergic population and 9 for the overall Dutch population.

Chapter 6

Concluding remarks

6.1 Food groups for allergen risk assessment

When the Ph.D. started they were no published food consumption data to use in food allergy risk assessment. As the amount of data derived from the National Food Consumption Surveys is huge, some work on organizing the data is needed before it can be used. Thus, the aim of the work was to organize the consumption data in relevant food groups that are easy to use.

The main challenge was that similar work on finding an automated procedure to create food groups for allergen risk assessment has never been done. So, the statistical investigations were not based on any already published method, as no similar ones were found in the bibliography.

Food groups for allergen risk assessment were created using the consumption data from Netherlands, France and Denmark jointly. The food groups created will serve as an input to the tiered risk assessment approach developed within the work package. From the food groups a point estimate can be calculated from the consumption distributions and used in the TIER 1 risk assessment. And the whole consumptions distributions can be used to estimate the chance of allergic

within different population in the TIER 2 risk assessment. it is now possible for a food producer or an authority to make food allergy risk assessment covering one country or the three countries.

The ultimate goal was to create a European food consumption database for food allergy risk assessment. Using the consumption data from three countries is the first attempt to create such a database. But, there is still some work left to fulfil this goal. The investigations made on the surveys design showed that the survey design have a small impact on the estimation of the of allergic reaction risk assessment, hence the food groups outline. So combining food consumption surveys with different designs is possible. Furthermore, as no subjective choice as been made for the criterion, it is possible to integrate consumption data from another country, as the procedure is designed in a way that it can be applied easily to any National Food consumption Survey in Europe.

Moreover, in article 1, it was highlighted that outline of the groups is similar for the three countries. However, only countries from Western Europe were included in the investigation. It can be expected that food groups made with consumption from another country will have similar outline. But, among the 28 countries in Europe, some countries might have really different consumption patterns than the one used in the study. Then, if necessary, an additional level can be included in the procedure , so the countries can also be grouped. Still keeping in mind that the homogeneity of food consumption patterns within each food group is the main concern, so we don't end up with a too high number of food groups.

6.2 Probabilistic risk assessment

In order to recommend the best approach for estimating the risk of allergic reaction, the two probabilistic methods that are usually used to estimate the risk of allergic reaction in different populations were compared. The two methods use different approaches: one uses Frequentist simulations and the other one uses combination of second order Monte-Carlo simulations and Bayesian inferences. No differences in the risk estimation was found between the two methods. However, it was discovered that one of the method that claimed to integrate uncertainty on the risk calculation actually calculated the simulation error. Thus, an alternative approach was suggested, so the uncertainty on the risk estimation could be estimated with second order Monte-Carlo simulations.

Furthermore, an example of how an uncertainty analysis could be run with the different approaches was detailed. Thus, the input parameters which add the most uncertainty on the risk estimation can be identified. When new data on the different inputs become available, it will be possible to do the same analysis and identify which improvement was made on the precision of the risk estimation. It can also be used to assess the quality of the concentration data for which the allergen risk assessment is performed, and then decide if the outcome can be trusted.

Some further investigations can be made on how to integrate back calculation in the probabilistic risk assessment: i.e. from a specific risk in the population that authorities do not want to exceed how much is the maximum level of allergen contamination within the food product.

Finally, all the investigations made on both food groups and the probabilistic risk assessment were joined in an application that makes the probabilistic risk assessment accessible for everyone. The open-source software R was selected to implement the probabilistic risk assessment, so it can be easily shared with the Work package partners. Furthermore, the R code can be updated when improvements on the method are found. The data included in the application can also be updated when new data are produced: either food consumption data from another country or threshold values for another allergen.

Appendix A

Survival modelling of challenge data with R: Frequentist and Bayesian comparison

The report presents how the threshold data can be fitted with a survival model using a Weibull distribution both Frequentist and Bayesian way using the R software.

Survival modeling of challenge data: frequentist and Bayesian comparison

Sophie BIROT
DTU Compute, Statistics and data analysis

January 15, 2016

1 Challenge data

The risk is calculated taking into account the threshold distribution of a specific allergen. The quantity needed to trigger an allergic reaction vary among the population and according to the allergen. Selected threshold data from the VITAL project were shared (data available in publications), this type of data is interval censored. As the exact dose which triggers an allergenic reaction cannot be measured precisely, 2 bounds are measured. The NOAEL (No Observable Adverse Effect Level) which is the level just before the allergic reaction occurs. And the LOAEL (Lowest Observed Adverse Effect Level) which is the lowest level which triggers allergic reaction.

```
peanutData <- read.csv("C:/Users/sobi/Desktop/Peanut data/peanutData.csv")
```

1.1 Survival analysis

2 Weibull link function

2.1 Frequentist modeling

```
WeibullReg <- survreg(Surv(time=cumNOAEL, time2=cumLOAEL, type="interval2")~1,data=peanutData,na.action=
```

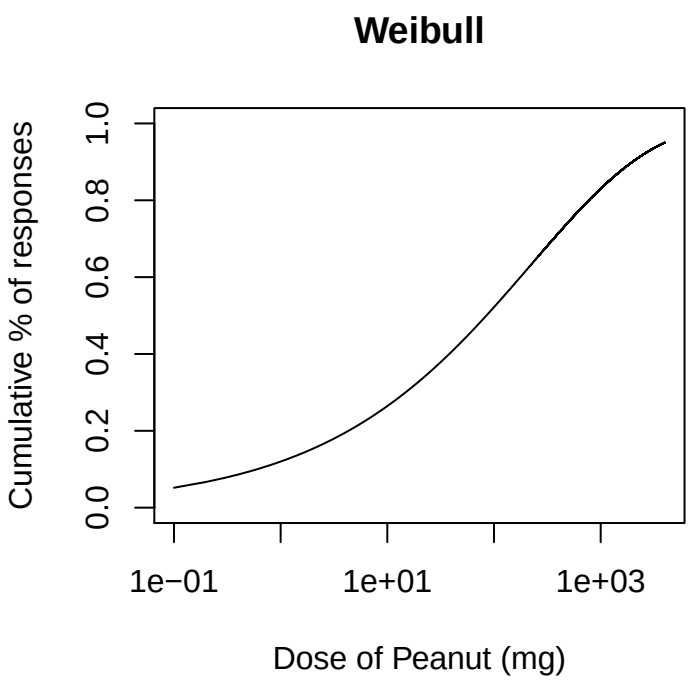
```
## survreg(formula = Surv(time = cumNOAEL, time2 = cumLOAEL, type = "interval2") ~
##     1, data = peanutData, na.action = na.omit, dist = "weibull")
```

	Value	Std. Error	z
(Intercept)	5.40	0.23	23.60
Log(scale)	0.97	0.07	13.59

Table 1: Survival analysis, peanut challenge data

So, the scale parameter was found to be 2.63 for the158 observations.

```
t <- seq(0.1,4000,0.1)
survx3=pweibull(t, scale = exp(WeibullReg$coeff), shape = 1/WeibullReg$scale)
plot(t,survx3,type='l',xlab="Dose of Peanut (mg)",ylab=paste("Cumulative % of responses"),ylim=c(0,1),lc
```




```
WeibullFit=predict(WeibullReg , type='quantile',p=c(0.01,0.05,0.1,0.5),se=T)

#Calculating 95% confidence interval

weibull=cbind(round(unique(WeibullFit[[1]][,1]),3),round(unique(WeibullFit[[1]][,2:4]),2))
dose=unique(WeibullFit[[1]])
stde=unique(WeibullFit[[2]])
ldose=log(dose)
stderr=stde/dose
upper=cbind(round(exp(ldose+1.96*stderr)[,1],3),t(round(exp(ldose+1.96*stderr)[,2:4],2)))
lower=cbind(round(exp(ldose-1.96*stderr)[,1],3),t(round(exp(ldose-1.96*stderr)[,2:4],2)))

weibull=rbind(weibull, lower,upper)
rownames(weibull)=c("Estimate","Lower","Upper")
colnames(weibull)=c("ED01","ED05","ED10","ED50")
print(xtable(weibull,caption="Eliting Dose estimates and confidence interval for Weibull distribution, P
```

	ED01	ED05	ED10	ED50
Estimate	0.00	0.09	0.60	84.85
Lower	0.00	0.03	0.22	51.70
Upper	0.01	0.32	1.67	139.25

Table 2: Eliting Dose estimates and confidence interval for Weibull distribution, Peanut

2.2 Bayesian modeling

```
source("C:/Users/sobi/Desktop/DBDA2Eprograms/DBDA2E-utilities.R")
```

```
#Preparing the thresholds data for JAGS
yBin=rep(NA,dim(peanutData)[1])
threshMat=as.matrix(peanutData[,10:11])
yBin[rowSums(is.na(threshMat))==0]=1
yBin[peanutData[,13]==1]=0
threshMat[peanutData[,13]==1,1:2]=cbind(threshMat[peanutData[,13]==1,2],threshMat[peanutData[,13]==1,2]+
yBin[peanutData[,14]==1]=2
threshMat[peanutData[,14]==1,1:2]=cbind(threshMat[peanutData[,14]==1,1]-1,threshMat[peanutData[,14]==1,1
```

```
#Data list for JAGS
y=rep(NA,length(yBin))
dataList = list(
  y = y ,
  yBin = yBin ,
  threshMat = threshMat ,
  Ntotal = length(y)
)
```

```
#JAGS bayesian model
modelString = "
model {
for ( i in 1:Ntotal ) {
y[i] ~ dweib( nu , lambda )  # JAGS parameterization
yBin[i] ~ dinterval( y[i] , threshMat[i, ] )
}
scalereg<-1/a
nu <- a   # a is shape in R dweibull. Here assumed same for all groups.
a ~ dgamma( 0.001 , 0.001) #Non-informative prior
lambda<- 1/b^a  # b is scale in R dweibull. Different for each group.
intercept<-log(b)
b ~ dgamma( 0.001 , 0.001) #Non-informative prior
}
" # close quote for modelString
writeLines( modelString , con="TEMPmodel.txt" )
```

```
# Initialize the missing y values:
# intial values of censored data:
yInit = rep( NA , length(y) )
for ( i in 1:length(y) ) {
  if ( is.na(y[i]) ) { # if y is censored
    if ( yBin[i]==0 ) {
      yInit[i] = threshMat[i,1]/2
    } else if ( yBin[i]==ncol(threshMat) ) {
      yInit[i] = threshMat[i,ncol(threshMat)]+1
    } else {
      yInit[i] = (threshMat[i,yBin[i]]+threshMat[i,yBin[i]+1])/2
```

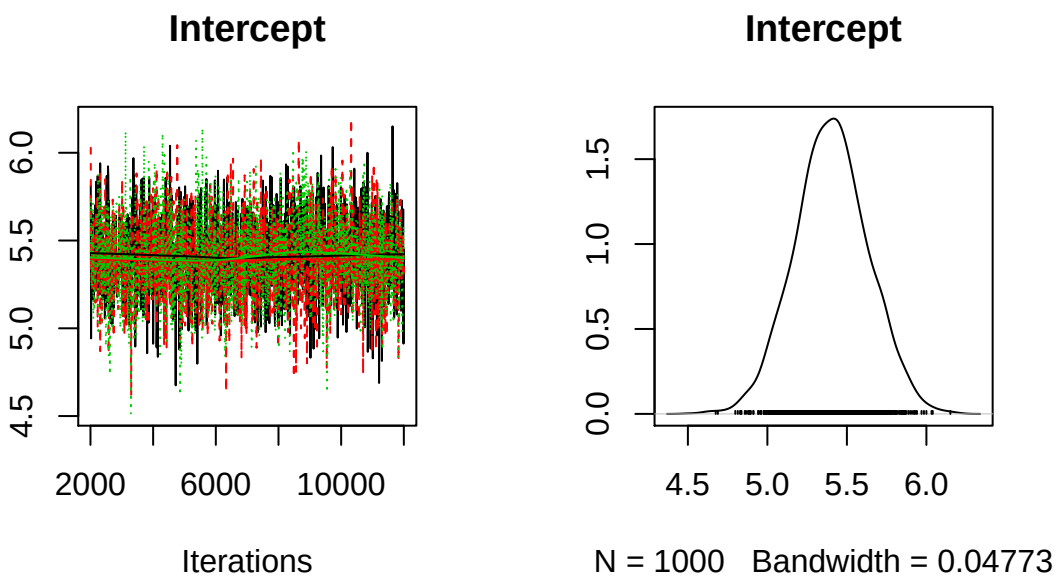
```
    }
  }
}
initsList = list( y=yInit )
```

```
#Run model on JAGS
library(rjags)
jagsModel = jags.model( file="TEMPmodel.txt" , data=dataList , inits=initsList , n.chains=3 , n.adapt=20

## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 158
##    Unobserved stochastic nodes: 160
##    Total graph size: 804
##
## Initializing model

#update( jagsModel , n.iter=2000 )
codaSamples = coda.samples( jagsModel , variable.names=c("nu","lambda","a","b","intercept","scalereg") ,
```

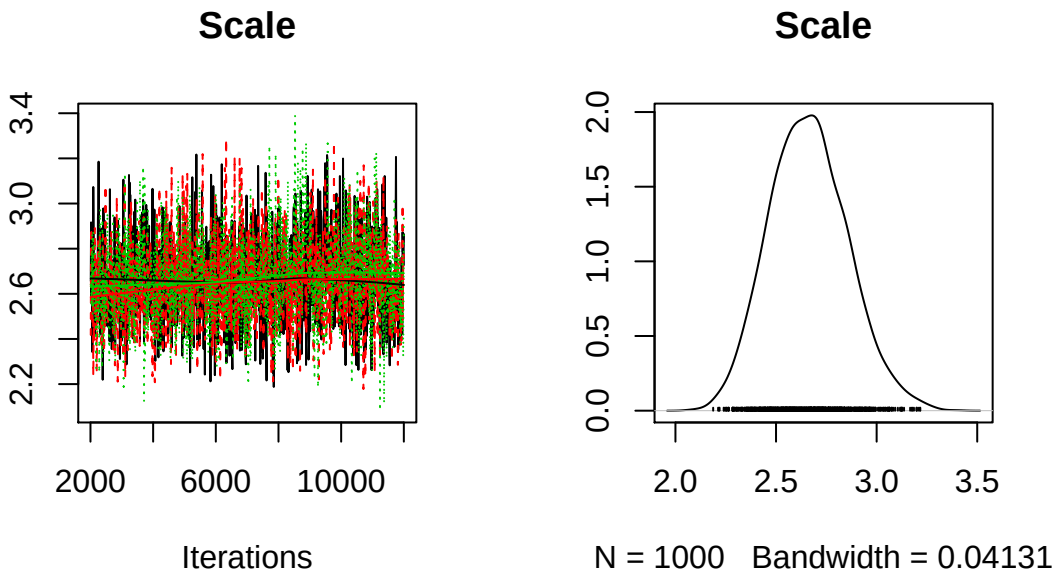
```
##
## Iterations = 2010:12000
## Thinning interval = 10
## Number of chains = 3
## Sample size per chain = 1000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##               Mean       SD  Naive SE Time-series SE
## a           0.3776  0.02737 0.0004996      0.0006458
## b         228.1379 53.44901 0.9758410      1.0638228
## intercept   5.4033  0.23103 0.0042180      0.0043051
## lambda      0.1321  0.02467 0.0004503      0.0004950
## nu          0.3776  0.02737 0.0004996      0.0006458
## scalereg    2.6621  0.19327 0.0035287      0.0047735
##
## 2. Quantiles for each variable:
##
##                2.5%      25%      50%      75%    97.5%
## a           0.32637   0.3581   0.3763   0.3961   0.4334
## b         142.55797 190.9253 221.8942 257.5313 350.1006
## intercept   4.95975   5.2519   5.4022   5.5511   5.8582
## lambda      0.08873   0.1148   0.1298   0.1479   0.1837
## nu          0.32637   0.3581   0.3763   0.3961   0.4334
## scalereg    2.30725   2.5244   2.6573   2.7922   3.0640
```

Diagonistics plot

```
plot(codaSamples[,"intercept"],main="Intercept")
```

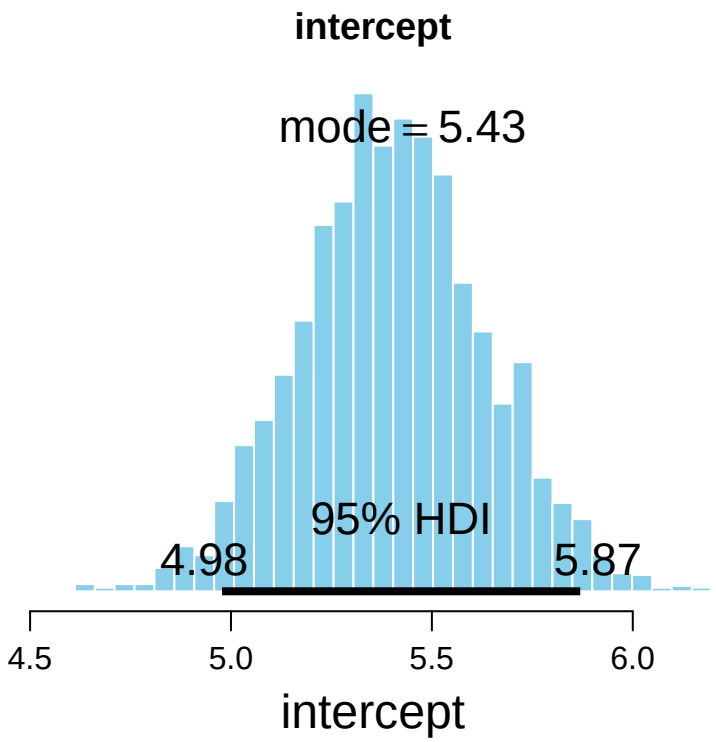



```
plot(codaSamples[,"scalereg"],main="Scale")
```

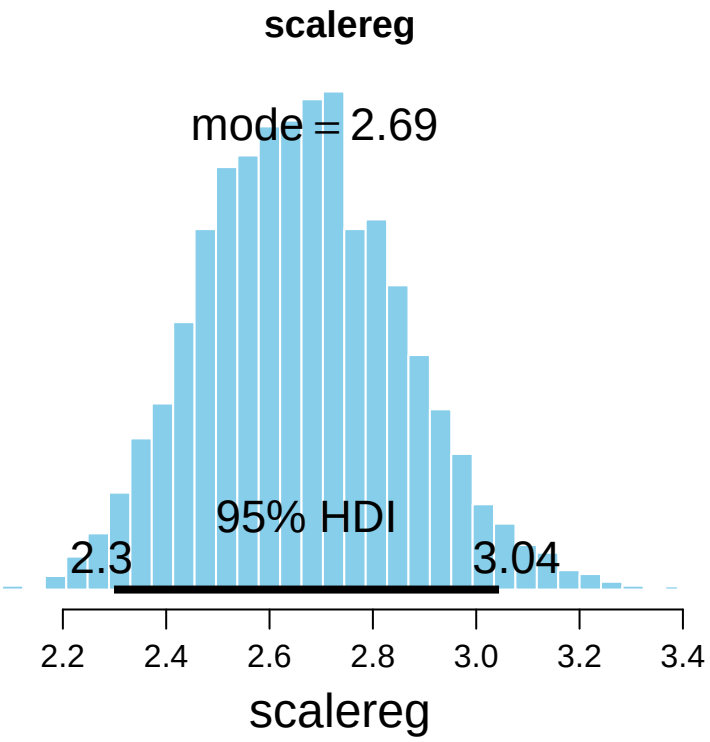



```
#saveGraph( file=paste0(fileNameRoot,"ThetaDiag") , type="eps" )
# Posterior descriptives:
openGraph(height=3,width=4)
```

```
par( mar=c(3.5,0.5,2.5,0.5) , mgp=c(2.25,0.7,0) )
plotPost( codaSamples[,"intercept"] , main="intercept" , xlab="intercept" )
```




```
##                ESS     mean   median     mode hdiMass   hdiLow  hdiHigh
## intercept 2462.909 5.403269 5.402201 5.426972    0.95 4.977775 5.869306
##           compVal pGtCompVal ROPElow ROPEhigh pLtROPE pInROPE pGtROPE
## intercept      NA         NA      NA       NA      NA      NA      NA

openGraph(height=3,width=4)
par( mar=c(3.5,0.5,2.5,0.5) , mgp=c(2.25,0.7,0) )
plotPost( codaSamples[,"scalereg"] , main="scalereg" , xlab="scalereg" )
```




```
##               ESS     mean   median     mode hdiMass  hdiLow  hdiHigh
## scalereg 1669.038 2.662129 2.657265 2.687371    0.95 2.29904 3.044183
##          compVal pGtCompVal ROPElow ROPEhigh pLtROPE pInROPE pGtROPE
## scalereg      NA         NA      NA       NA      NA      NA      NA
```

Appendix B

Probabilistic risk modelling: uncertainty and variability assessment

The report presents some investigations about the comparison of the methods used for allergen probabilistic risk assessment, the uncertainty and variability assessment for these methods and the impact of the number of iterations and replications on the risk uncertainty evaluation.

Probabilistic risk modelling: uncertainty and variability assessment

Sophie BIROT, Per B. Brockhoff
DTU Compute, Statistics and data analysis

March 22, 2016

1 Introduction

We have the three distributions:

$$\text{Consumption:} X \sim F_x$$

$$\text{Contamination:} Y \sim F_y$$

$$\text{Threshold:} Z \sim F_z$$

$$\text{Allergy outcome:} U = 1, \text{ if } Z < XY, \text{ and 0 otherwise}$$

$$U \sim \text{bernoulli}(p) \text{ , where } p_u = P(Z < XY)$$

And formally it is assumed that the three random variables X, Y and Z are independent.

The probability p_u is mathematically a function of the three distributions:

$$p_u = f(F_x, F_y, F_z)$$

If these three distributions are given by parameters, then this a function of these:

$$p_u = f(\theta_x, \theta_y, \theta_z)$$

E.g. they could be three log-normal distributions:

$$p_u = f(\mu_x^L, \sigma_x^L, \mu_y^L, \sigma_y^L, \mu_z^L, \sigma_z^L)$$

In fact, in this case it becomes:

$$p_u = 1 - \Phi\left(\frac{\mu_z^L - \mu_x^L - \mu_y^L}{\sqrt{(\sigma_x^L)^2 + (\sigma_y^L)^2 + (\sigma_z^L)^2}}\right)$$

But generally the form of this can be more complicated and without an analytic expression like this.

2 Input variables of the allergen risk model

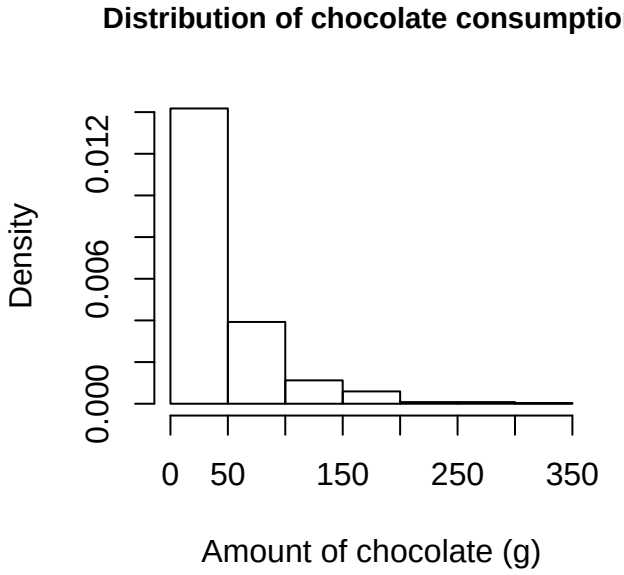
2.1 Q: consumption data

As the Rimbaud article is focused on contamination of chocolate product, we will use the consumption data from the French National Food Consumption and the 19C food group (chocolate and chocolate products). The largest food consumption of all the food items as the risk assessment is design to protect the consumers from an acute reaction.

N=1284 consumers over the 2624 participants in the consumption survey were found to consume chocolate.

Mean	SD	2.5%	50%	97.5%
47.48	49.98	3.00	30.00	200.00

Table 1: Empirical distribution of chocolate consumption (g)



2.2 p: probability of allergen presence

Posterior distribution is calculated analytically:

$$p \sim Beta(1 + x, 1 + n_1 - x) \tag{1}$$

with x(=100) the number of contaminated samples and n_1(=275) the total number of analyzed samples:

$$p \sim Beta(101, 176) \tag{2}$$

The parameters p is simulated n=10^4 times, as described in the paper.

```
p=rbeta(n,101,176)
```

Mean	SD	2.5%	50%	97.5%
0.36	0.03	0.31	0.36	0.42

Table 2: Probability of the allergen presence's distribution

2.3 λ and c_j: allergen concentration levels

The concentration of allergens int he contaminated sample j=1,...,J=100, noted as $c_j$is modeled with an exponential distribution:

$$c_j|\lambda \sim Exp(\lambda) \tag{3}$$

The conjugate gamma distribution $Gamma(\alpha, \beta)$ is used for the prior distribution of the paramater λ, with vague prior parmater $\alpha = \beta = 10^{-3}$ The posterior distribution of of the parameter λ can be calculated analytically:

$$\lambda|c \sim Gamma(\alpha + J_1, \beta + \sum_{j=1}^{J} c_j) \tag{4}$$

The parameters of λ's distribution are estimated based on the number of noncensored data J_1=53, and the sum of the peanut protein concentration values $\sum_{j=1}^{J} c_j$=419.81:

$$\lambda|c \sim Gamma(53 + 420) \tag{5}$$

```
lambda=rgamma(n, 53, rate = 419.81)
```

Mean	SD	2.5%	50%	97.5%
0.13	0.02	0.09	0.13	0.16

Table 3: Parameter of the allergen concentration's distribution

As describe above, $C \sim Exp(\lambda)$, so we can calculate the concentration distribution for selected value from the λ distribution: 2.5%, 50% and 97.5% percentile.

```
test1=rexp(n,rate=median(lambda))
```

Mean	SD	2.5%	50%	97.5%
8.02	7.99	0.21	5.57	28.97

Table 4: Concentration's distribution, median

```
test2=rexp(n,rate=quantile(lambda,probs=0.025))
```

```
test3=rexp(n,rate=quantile(lambda,probs=0.975))
```

2.4 a and b: parameters of the dose response equation

The probability of an allergic person reacting to the dose d_l is defined by the Weibull cumulative distribution function:

$$DR(d) = 1 - e^{-(\frac{d_l}{b})^a} \tag{6}$$

The data used are the one gather in the VITAL project and shared by the partners.

Mean	SD	2.5%	50%	97.5%
10.59	10.73	0.27	7.30	39.01

Table 5: Concentration's distribution, 2.5 percentile

Mean	SD	2.5%	50%	97.5%
6.11	6.05	0.16	4.28	22.43

Table 6: Concentration's distribution, 2.5 percentile

```
peanutData <- read.csv("C:/Users/sobi/Desktop/Peanut data/peanutData.csv")
```

```
source("C:/Users/sobi/Desktop/DBDA2Eprograms/DBDA2E-utilities.R")
```

```
#JAGS bayesian model
modelString = "
model {
for ( i in 1:Ntotal ) {
y[i] ~ dweib( nu , lambda )  # JAGS parameterization
yBin[i] ~ dinterval( y[i] , threshMat[i, ] )
}
scalereg<-1/a
nu <- a   # a is shape in R dweibull. Here assumed same for all groups.
a ~ dgamma( 0.001 , 0.001) #Non-informative prior
lambda<- 1/b^a  # b is scale in R dweibull. Different for each group.
intercept<-log(b)
b ~ dgamma( 0.001 , 0.001) #Non-informative prior
}
" # close quote for modelString
writeLines( modelString , con="TEMPmodel.txt" )
```

```
#Run model on JAGS
library(rjags)
jagsModel = jags.model( file="TEMPmodel.txt" , data=dataList , inits=initsList , n.chains=2 , n.adapt=20

## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 158
##    Unobserved stochastic nodes: 160
##    Total graph size: 804
##
## Initializing model

#update( jagsModel , n.iter=2000 )
codaSamples = coda.samples( jagsModel , variable.names=c("a","b") ,n.iter=10000,thin=1 )
chain=codaSamples[[1]]
```

```
summary(window(codaSamples))

##
## Iterations = 2001:12000
## Thinning interval = 1
## Number of chains = 2
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##      Mean       SD  Naive SE Time-series SE
## a   0.376  0.02698 0.0001908      0.0006582
## b 228.320 54.67575 0.3866160      0.8166171
##
## 2. Quantiles for each variable:
##
##      2.5%      25%      50%      75%  97.5%
## a   0.324   0.3577   0.3757   0.3938   0.43
## b 139.991 189.6462 221.9848 259.7463 353.22
```

	Mean	SD	2.5%	50%	97.5%
a	0.38	0.03	0.32	0.38	0.43
b	228.32	54.68	139.99	221.98	353.22

Table 7: Parameters of the dose response's distribution

3 No uncertainty

3.1 Risk expression

In Tier 1 we compute p_u by plugging in our best estimate of the three distributions:

$$\hat{p}_u = f(\hat{F}_x, \hat{F}_y, \hat{F}_z)$$

Most often we use a parametric version of these:

$$\hat{p}_u = f(\hat{\theta}_x, \hat{\theta}_y, \hat{\theta}_z)$$

where each θ typically would be a mean and a variance (or similarly) of the distributions. As indicated, for the triple log-normal model this number can be calculated explicitly by the standard model distribution function.

3.2 Risk computation

```
# Consumption: from data (kg)

m1 = mean(FR.choc[, 2]) * 10^-3
s1 = sd(FR.choc[, 2]) * 10^-3

mm1 = log((m1^2)/sqrt(m1^2 + s1^2))
ss1 = sqrt(log(1 + (s1/m1)^2))
```

```
# Threshold distribution Mean from Weibull
# distribution with estimated parameters from
# Bayesian modelling
amean = summary(window(codaSamples))[[1]][which(rownames(summary(window(codaSamples))[[1]][,
    1:2]) == "a"), 1]
bmean = summary(window(codaSamples))[[1]][which(rownames(summary(window(codaSamples))[[1]][,
    1:2]) == "b"), 1]

# alpha <- bmean*gamma(1+1/amean) beta <-
# sqrt(bmean^2*gamma(1+2/amean)-(gamma(1+1/amean))^2)
```

```
# Concentration distribution lambda mean for
# exponential distribution
lambdamean = mean(lambda)
```

Generally, in this world it is said that the p_u is found by simulation which is also fine, as long as N is large enough to exclude any simulation error:

```
# The "simulation based"  computation of p_u:
NbSim <- 1000000

x1 <- rlnorm(NbSim , mm1, ss1)
#x2 <- rlnorm(NbSim , mm2, ss2)
x2 <- rexp(NbSim , lambdamean)
#x3 <- rlnorm(NbSim , mmz, ssz)
x3 <- rweibull(NbSim , amean, bmean)

## And counting again:

mean(x1*x2>x3)

## [1] 0.069653
```

Redoing such simulations several times to find and show the simulation error is irrelevant in an application context and not necessary - it may be estimated by binomial probability theory:

$$SE_{\hat{p}_u}^{\text{sim}} = \sqrt{\frac{\hat{p}_u(1-\hat{p}_u)}{N}} = \sqrt{\frac{7\%(1-7\%)}{10^6}} = 2.55 \times 10^{-4}$$

And for practical use, N should be chosen large enough to make this error negligable, and it should NOT be communicated to anyone, as it can only induce confusion.

4 With uncertainty (TIER 2)

This is an investigation of how the (sampling) uncertainty of all the distributions propagates through the non-linear $\hat{p}_u$ computation. With sampling distributions for ALL parameters (ALSO variances, right?) we can investigate this, actually for each distribution separately of for all of them jointly. Formally, it amounts to considering the $\hat{p}_u$ as a random variable as a function of the data (in the usual way, when talking about smapling statistics)

$$\hat{P}_u^X = f(\hat{\Theta}_X, \hat{\theta}_y, \hat{\theta}_z)$$

where $\hat{\Theta}_X$ is a random sampling statistics (to investigate the ucertainty induced by consumption data sampling), and the other two ones are held at the observed estimated parameters). Or:

$$\hat{P}_u^Y = f(\hat{\theta}_x, \hat{\Theta}_Y, \hat{\theta}_z)$$

Or:

$$\hat{P}_u^Z = f(\hat{\theta}_x, \hat{\theta}_y, \hat{\Theta}_Z)$$

Or all of them (which is what people here usually do:)

$$\hat{P}_u = f(\hat{\Theta}_X, \hat{\Theta}_Y, \hat{\Theta}_Z)$$

It raises the following question to Ben and Joe: When they say they sample "the means" from the sampling distributions (once for each 25): What about the variances for the 3 distributions - how do they chose them? They should be used in each risk calculation, I would say?

In practice this is what we have done so far, I think: For a high number of times, M, (NOT only 25 nor 100), select parameter values from the sampling distributions, and for each selection make the almost exact (large N) $\hat{P}_u$ computation each time. In this, way we can by looking at this distribution of $\hat{P}_u$ show the impact of uncertainty.

4.1 FARRP risk calculation

```
K=nbRep[1] #Number of sets of parameters

a=sample(chain[,1],K)
b=sample(chain[,2],K)
l=sample(lambda,K)
#P=sample(p,K)
P=rep(1,K)
```

```
N2=10000
#N2=NbSim
occurence2=sapply(P,function(x) rbinom(N2, 1, x))
```

```
concentration2=sapply(l,function(x) rexp(N2, x))
```

```
consumption2=replicate(K,sample(FR.choc[,2],N2,replace=T))

#apply(consumption,2,mean)

exposure2= occurence2*concentration2*consumption2*10^-3
#apply(exposure,2,mean)
```

```
risk2=function (a,b,N){
  R=rweibull(N,shape=a,scale=b)
}
R2=mapply(risk2,a,b,N2)
```

25	100	1000	10000
6.84	6.89	7.07	7.02

Table 8: Mean of Risk of Allergic reaction for K replications

25	100	1000	10000
1.31	1.49	1.40	1.40

Table 9: Standard deviation of Risk of Allergic reaction for K replications

Some analysis of the 25-procedure of Joe and Ben and explanation of my comments in the email: I simply look at this as a regular $n = 25$ one-sample statistics setting: and the outcome is $\bar{x} = 8.7$ and $s = 3.1$:

$$\hat{\text{Var}}(\bar{X}) = \frac{s}{\sqrt{25}} = \frac{3.1}{5} = 0.6$$

Now, for the example above:

25	100	1000	10000
0.26	0.15	0.04	0.01

Table 10: Variance of Risk of Allergic reaction for K replications

And then I did a mistake in my Coefficient of Variation remark, which is instead $0.6/8.7 = 7\%$

25	100	1000	10000
3.83	2.17	0.62	0.20

Table 11: Coefficient of Variation of Risk of Allergic reaction for K replications

And is is wellknown standard teory that the relative standard error of the s^2 is $\sqrt{1/(n-1)}$ - comes from the variance of the $\chi^2(n-1)$-distribution which is the sampling distribution of

$$\frac{(n-1)s^2}{\sigma^2}$$

A $\sqrt{(1/24)} = 0.20 = 20\%$ relative error on the variance hence becomes a 14% relative simulation error on the SD.

25	100	1000	10000
20.41	10.05	3.16	1.00

Table 12: Relative error on the variance of Risk of Allergic reaction for K replications

4.2 FARRP risk calculation: uncertainty on concentration parameters only

```
sysTime[1]=system.time({
K=nbRep[1] #Number of sets of parameters

a=sample(chain[,1],K)
b=sample(chain[,2],K)
l=sample(lambda,K)
#P=sample(p,K)
P=rep(1,K)
N2=10000
#N2=NbSim
occurence2=sapply(P,function(x) rbinom(N2, 1, x))

concentration2=sapply(l,function(x) rexp(N2, x))
```

```
#concentration2=replicate(K,rexp(N2,lambdamean))



#consumption2=replicate(K,sample(FR.choc[,2],N2,replace=T))
consumption2=replicate(K,rlnorm(N2,mm1,ss1))
#apply(consumption,2,mean)

exposure2= occurence2*concentration2*consumption2
#apply(exposure,2,mean)

# risk2=function (a,b,N){
#   R=rweibull(N,shape=a,scale=b)
# }
# R2=mapply(risk2,a,b,N2)
R2=replicate(K,rweibull(N2,shape=amean,scale=bmean))
risk=colSums(exposure2>R2)/N2
meanRisk[1]=mean(risk)
sdRisk[1]=sd(risk)
})[1]
```

25	100	1000	10000
7.10	6.99	6.98	7.02

Table 13: Mean of Risk of Allergic reaction for K replications

25	100	1000	10000
0.38	0.45	0.43	0.43

Table 14: Standard deviation of Risk of Allergic reaction for K replications

Some analysis of the 25-procedure of Joe and Ben and explanation of my comments in the email: I simply look at this as a regular $n = 25$ one-sample statistics setting: and the outcome is $\bar{x} = 8.7$ and $s = 3.1$:

$$\hat{\text{Var}}(\bar{X}) = \frac{s}{\sqrt{25}} = \frac{3.1}{5} = 0.6$$

Now, for the example above:

25	100	1000	10000
0.08	0.04	0.01	0.00

Table 15: Variance of Risk of Allergic reaction for K replications

And then I did a mistake in my Coefficient of Variation remark, which is instead $0.6/8.7 = 7\%$

25	100	1000	10000
1.08	0.64	0.20	0.06

Table 16: Coefficient of Variation of Risk of Allergic reaction for K replications

And is is wellknown standard teory that the relative standard error of the s^2 is $\sqrt{1/(n-1)}$ - comes

from the variance of the $\chi^2(n-1)$-distribution which is the sampling distribution of

$$\frac{(n-1)s^2}{\sigma^2}$$

A $\sqrt{(1/24)} = 0.20 = 20\%$ relative error on the variance hence becomes a 14% relative simulation error on the SD.

25	100	1000	10000
20.41	10.05	3.16	1.00

Table 17: Relative error on the variance of Risk of Allergic reaction for K replications

25	100	1000	10000
0.13	0.48	4.64	47.02

Table 18: User system time for risk computation (s)

4.3 FARRP risk calculation: uncertainty on consumption parameters only

```
sysTime[1]=system.time({
K=nbRep[1] #Number of sets of parameters

a=sample(chain[,1],K)
b=sample(chain[,2],K)
l=sample(lambda,K)
#P=sample(p,K)
P=rep(1,K)
N2=10000
#N2=NbSim
occurence2=sapply(P,function(x) rbinom(N2, 1, x))

#concentration2=sapply(l,function(x) rexp(N2, x))
concentration2=replicate(K,rexp(N2,lambdamean))



consumption2=replicate(K,sample(FR.choc[,2],N2,replace=T))
#consumption2=replicate(K,rlnorm(N2,mm1,ss1))
#apply(consumption,2,mean)

exposure2= occurence2*concentration2*consumption2*10^-3
#apply(exposure,2,mean)

risk2=function (a,b,N){
  R=rweibull(N,shape=a,scale=b)
}
R2=mapply(risk2,a,b,N2)
risk=colSums(exposure2>R2)/N2
meanRisk[1]=mean(risk)
sdRisk[1]=sd(risk)
})[1]
```

25	100	1000	10000
6.85	7.18	7.02	6.99

Table 19: Mean of Risk of Allergic reaction for K replications

25	100	1000	10000
1.36	1.31	1.36	1.37

Table 20: Standard deviation of Risk of Allergic reaction for K replications

Some analysis of the 25-procedure of Joe and Ben and explanation of my comments in the email: I simply look at this as a regular $n = 25$ one-sample statistics setting: and the outcome is $\bar{x} = 8.7$ and $s = 3.1$:

$$\hat{\text{Var}}(\bar{X}) = \frac{s}{\sqrt{25}} = \frac{3.1}{5} = 0.6$$

Now, for the example above:

25	100	1000	10000
0.27	0.13	0.04	0.01

Table 21: Variance of Risk of Allergic reaction for K replications

And then I did a mistake in my Coefficient of Variation remark, which is instead $0.6/8.7 = 7\%$

25	100	1000	10000
3.96	1.82	0.61	0.20

Table 22: Coefficient of Variation of Risk of Allergic reaction for K replications

And is is wellknown standard teory that the relative standard error of the s^2 is $\sqrt{1/(n-1)}$ - comes from the variance of the $\chi^2(n-1)$-distribution which is the sampling distribution of

$$\frac{(n-1)s^2}{\sigma^2}$$

A $\sqrt{(1/24)} = 0.20 = 20\%$ relative error on the variance hence becomes a 14% relative simulation error on the SD.

25	100	1000	10000
20.41	10.05	3.16	1.00

Table 23: Relative error on the variance of Risk of Allergic reaction for K replications

25	100	1000	10000
0.10	0.34	3.41	34.10

Table 24: User system time for risk computation (s)

4.4 FARRP risk calculation: uncertainty on threshold parameters only

```
sysTime[1]=system.time({
K=nbRep[1] #Number of sets of parameters

a=sample(chain[,1],K)
b=sample(chain[,2],K)
l=sample(lambda,K)
#P=sample(p,K)
P=rep(1,K)
N2=10000
#N2=NbSim
occurence2=sapply(P,function(x) rbinom(N2, 1, x))

#concentration2=sapply(l,function(x) rexp(N2, x))
concentration2=replicate(K,rexp(N2,lambdamean))



#consumption2=replicate(K,sample(FR.choc[,2],N2,replace=T))
consumption2=replicate(K,rlnorm(N2,mm1,ss1))
#apply(consumption,2,mean)

exposure2= occurence2*concentration2*consumption2
#apply(exposure,2,mean)

risk2=function (a,b,N){
  R=rweibull(N,shape=a,scale=b)
}
R2=mapply(risk2,a,b,N2)
risk=colSums(exposure2>R2)/N2
meanRisk[1]=mean(risk)
sdRisk[1]=sd(risk)
})[1]
```

25	100	1000	10000
6.98	7.17	7.20	7.22

Table 25: Mean of Risk of Allergic reaction for K replications

25	100	1000	10000
1.30	1.42	1.35	1.39

Table 26: Standard deviation of Risk of Allergic reaction for K replications

Some analysis of the 25-procedure of Joe and Ben and explanation of my comments in the email: I simply look at this as a regular $n = 25$ one-sample statistics setting: and the outcome is $\bar{x} = 8.7$ and $s = 3.1$:

$$\hat{\text{Var}}(\bar{X}) = \frac{s}{\sqrt{25}} = \frac{3.1}{5} = 0.6$$

Now, for the example above:

25	100	1000	10000
0.26	0.14	0.04	0.01

Table 27: Variance of Risk of Allergic reaction for K replications

And then I did a mistake in my Coefficient of Variation remark, which is instead $0.6/8.7 = 7\%$

25	100	1000	10000
3.74	1.99	0.59	0.19

Table 28: Coefficient of Variation of Risk of Allergic reaction for K replications

And is is wellknown standard teory that the relative standard error of the s^2 is $\sqrt{1/(n-1)}$ - comes from the variance of the $\chi^2(n-1)$-distribution which is the sampling distribution of

$$\frac{(n-1)s^2}{\sigma^2}$$

A $\sqrt{(1/24)} = 0.20 = 20\%$ relative error on the variance hence becomes a 14% relative simulation error on the SD.

25	100	1000	10000
20.41	10.05	3.16	1.00

Table 29: Relative error on the variance of Risk of Allergic reaction for K replications

25	100	1000	10000
0.13	0.46	4.57	46.36

Table 30: User system time for risk computation (s)

Appendix C

R code for the Shiny application

C.1 R code: launch shiny application

```
library(shiny)
path = setwd("C:/Users/sobi/Desktop/Shiny App/Risk_App")
runApp(path)
```

C.2 R code: UI (interface) shiny application

```
shinyUI(fluidPage(
  useShinyjs(),
  titlePanel(  title=div(img(src="kunlogo.png"), "Risk Assessment
   Application")),
  tabsetPanel(
    tabPanel("Risk Assessment",
             sidebarLayout(
```

```
                sidebarPanel(
                  wellPanel(h4("Challenge data"),
                            fileInput('file1', 'Choose CSV File',
                                      accept=c('text/csv', 'text/comma
                                      -separated-values,text/plain', '.csv')),
                            #tags$hr(),
                            checkboxInput('header', 'Header', TRUE),
                            radioButtons('sep', 'Separator',
                                         c(Comma=',',
                                           Semicolon=';',
                                           Tab='\t'),
                                         'Comma'),
                            radioButtons('quote', 'Quote',
                                         c(None='',
                                           'Double Quote'='"',
                                           'Single Quote'="'"),
                                         'Double Quote'),
                            actionButton('checkFile','Validate file'),

                            hidden(selectizeInput('Allergen','Select
                            the allergen variable',choices = "Pending upload" ,options
                            hidden(selectizeInput('Population','Select
                            the population variable',choices = "Pending upload",optio
                            hidden(selectizeInput('NOAEL','Select
                            the NOAEL variable',choices = "Pending upload",options = l
                            hidden(selectizeInput('LOAEL','Select
                            the LOAEL variable',choices = "Pending upload",options = l
                            hidden(actionButton('checkVariable',
                            'Validate variables')),
                            hidden(selectizeInput('SelectAllergen',
                            'Select the allergen:',choices = "Pending upload",options
                            uiOutput('boxPopulation')
                  ),

                  hidden(wellPanel(h4("Consumption data"),id = "consumption",
                                   selectizeInput(
                                     'Country', 'Select the countries combined or a sp
                                     options = list(
                                       placeholder = 'Please select an option below',
                                       onInitialize = I('function() { this.setValue("
                                     )
                                   ),
                                   selectizeInput(
```

```
                                'Population', 'Select a population:', choices
                                options = list(
                                  placeholder = 'Please select an option belo
                                  onInitialize = I('function() { this.setValu
                                )
                              ),
                              uiOutput('boxGroup'),
                              uiOutput('boxFoodItm'))),
                  hidden(wellPanel(h4("Contamination data"),id = "contamination",
                              textInput("meanCont",label="Mean for the level
                              textInput("sdCont",label="SD for the level of c
                              textInput("NbCont",label="Number of points for
                              textInput("meanChanceCont",label="Mean for the
                              h5(helpText("Select below if you would like to
                              checkboxInput(inputId = "sdChanceContBox", "SD
                              hidden(textInput("sdChanceCont",label="SD for t
                              hidden(textInput("NbChanceCont",label="Number o
                              #,actionButton('update','Calculate')
                  )),
                  hidden(wellPanel(id = "submit",hidden(actionButton('update','Cal
                ),
                mainPanel(tableOutput('contents'),
                          hidden(wellPanel(style = "background-color: #ffffff;",id
                                           h2("Populations Risks' estimation"),
                                           br(),
                                           #plotOutput('graph'),
                                           h3(textOutput("title1")),
                                           h3(textOutput("title2")),
                                           hr(),
                                           h4("Mean Consumption (g)"),
                                           tableOutput("summaryCons"),
                                           h4("Chance of Consumption (%)"),
                                           tableOutput("summaryChance"),
                                           hr(),
                                           h3(textOutput("title3")),
                                           h3(textOutput("title4")),
                                           br(),
                                           hr(),
                                           h4("Gender = All"),
                                           tableOutput("RiskTable"),
                                           hr(),
                                           h4("Gender = Male"),
                                           tableOutput("RiskTableMale"),
```

```
                                        hr(),
                                        h4("Gender = Female"),
                                        tableOutput("RiskTableFemale"),
                                        hidden(downloadButton('downloadData','Downl
                                        hr(),
                                        br(),
                                        h2("Reactions' simulation plot (Log-Normal)
                                        plotOutput('reactionPlot')
                        ))))),
    tabPanel("Number of allergic reaction",
             sidebarLayout(
               sidebarPanel(
                 textInput("MSproducts",label="Product market share within the food g
                 textInput("percProducts",label="Number of products/packages/lots imp
                 actionButton('updateMS','Calculate')
               ),
               mainPanel(
                 h3("Number of allergic reaction predicted"),
                 tableOutput("RiskMS")
               ))))
```

C.3 R code: Server shiny application

```
shinyServer(function(input, output,
    session) {
    # Challenge data Show and hide
    # elements after the file is
    # uploaded and validate
    observeEvent(input$checkFile,
        {
            hide("header")
            hide("sep")
            hide("quote")
            hide("checkFile")
            hide("file1")
            show("Allergen")
            show("Population")
            show("NOAEL")
            show("LOAEL")
            show("checkVariable")
```

```
    })
# Show and hide elements after
# the variables are selected
# from the file
observeEvent(input$checkVariable,
    {
        hide("Allergen")
        hide("Population")
        hide("NOAEL")
        hide("LOAEL")
        hide("checkVariable")
        show("SelectAllergen")
        show("SelectPopulation")
        show("update")
        show("consumption")
        show("contamination")
        show("submit")
    })

# Change button status after a
# file is uploaded
observe({
    shinyjs::toggleState("checkFile",
        !is.null(dataChallenge()))
})

# Reactive dataset for
# challenge data A default file
# is uploaded or the user can
# select a file
dataChallenge <- reactive({
    inFile <- input$file1
    re1 <- reactive({
        is.null(inFile)
    })
    if (re1()) {
        read.csv("data/challenge.csv")
    } else {
        read.csv(inFile$datapath,
            header = input$header,
            sep = input$sep,
            quote = input$quote)
    }
```

```
    })
    # Show the uploaded file in a
    # table so the user can check
    # if the file is corerclty
    # uploaded
    output$contents <- renderTable({
        if (input$checkFile ==
            0) {
            dataChallenge()
        }
    })
    # Select allergen's variable
    # from the variables in the
    # file Allergen's variable
    # selected by default
    observe({
        updateSelectInput(session,
            "Allergen", choices = names(dataChallenge()),
            selected = "Allergen")
    })
    # Select population's variable
    # from the variables in the
    # file Population's variable
    # selected by default
    observe({
        updateSelectInput(session,
            "Population", choices = names(dataChallenge()),
            selected = "Population")
    })
    # Select NOAEL's variable from
    # the variables in the file
    # NOAEL's variable selected by
    # default
    observe({
        updateSelectInput(session,
            "NOAEL", choices = names(dataChallenge()),
            selected = "cumNOAEL")
    })
    # Select LOAEL's variable from
    # the variables in the file
    # LOAEL's variable selected by
    # default
    observe({
```

```
        updateSelectInput(session,
            "LOAEL", choices = names(dataChallenge()),
            selected = "cumLOAEL")
    })
    # Change button status when the
    # 4 variables are selected
    observe({
        shinyjs::toggleState("checkVariable",
            input$Allergen != "" &&
                input$Population !=
                  "" && input$NOAEL !=
                "" && input$LOAEL !=
                "")
    })
    # Upload the allergen's choices
    # from the selected allergen
    # variable
    observeEvent(input$checkVariable,
        {
            updateSelectInput(session,
                "SelectAllergen",
                choices = c(t(unique(dplyr::select_(dataChallenge(),
                  input$Allergen)))))
        })
    # Upload the population's
    # choices from the selected
    # population variables Only the
    # population available for the
    # selected allergen are
    # displayed
    observeEvent(input$checkVariable,
        {
            output$boxPopulation = renderUI(if (input$SelectAllergen ==
                "") {
                return()
            } else selectInput("SelectPopulation",
                "Select the population:",
                choices = c("All",
                  c(t(unique(dplyr::select_(dataChallenge()[dataChallenge()[,
                    names(dataChallenge()) ==
                      input$Allergen] ==
                    input$SelectAllergen,
                    ], input$Population)))))))
```

```
    })

# Update label's button to
# 'update'
observeEvent(input$update,
    {
        updateSelectInput(session,
            "update", label = "Update")
    })

# When 'update' button is
# clicked, the challenge data
# are filtered with the
# selected population and
# allergen


# Consumption data Return the
# dataset for the selected
# country
datasetInput <- eventReactive(input$Country !=
    "", {
    a = dataBaseConsumption[dataBaseConsumption$Country ==
        input$Country, ]
    a = a[, c(T, colSums(a[-1]) !=
        0)]
    # dataBaseConsumption%>%dplyr::filter(paste0('Country==',input$Country))
    return(a)
})
# Show the update button when
# input are not null
observeEvent(input$Country !=
    "" && input$Population !=
    "", {
    show("update")
})
# Select the food group among
# the ones available for the
# selected country
observeEvent(input$Country !=
    "" && input$Population !=
    "", {
    output$boxGroup = renderUI(if (input$Country ==
```

```
            "") {
            return()
        } else if (input$Country ==
            "Combined") {
            selectizeInput("SelectGroup",
                "Select a food group:",
                choices = c(do.call(paste,
                  c(unique(iFAAMgroups[iFAAMgroups$Country ==
                    "Combined",
                    1:2])[order(unique(iFAAMgroups[iFAAMgroups$Country ==
                    "Combined",
                    1:2])[, 1]),
                    ], sep = ": "))),
                options = list(placeholder = "Please select an option below",
                  onInitialize = I("function() { this.setValue(\"\"); }")))
        } else selectizeInput("SelectGroup",
            "Select a food group:",
            choices = c(do.call(paste,
                c(unique(iFAAMgroups[iFAAMgroups$FoodEx.Code %in%
                  colnames(datasetInput())[-1],
                  1:2])[order(unique(iFAAMgroups[iFAAMgroups$FoodEx.Code %in%
                  colnames(datasetInput())[-1],
                  1:2])[, 1]),
                  ], sep = ": "))),
            options = list(placeholder = "Please select an option below",
                onInitialize = I("function() { this.setValue(\"\"); }"))))
    })

    # Subset the iFAAM groups files
    # with the available food items
    # from the selected group
    selectGroup <- eventReactive(input$SelectGroup,
        {
            a = iFAAMgroups[iFAAMgroups$iFAAM.Group ==
                strsplit(input$SelectGroup,
                  ": ")[[1]][1],
                ]
            return(a)
        })

    # Show the food items common to
    # the iFAAM group selected and
    # the consumption data base of
```

```
    # the selected country
    observeEvent(input$SelectGroup,
        {
            output$boxFoodItm = renderUI(if (input$SelectGroup ==
                "") {
                return()
            } else {
                selectInput("SelectItm",
                  "Select a food item:",
                  choices = c("Group",
                    c(do.call(paste,
                      c(unique(selectGroup()[selectGroup()[,
                        5] %in%
                        colnames(dataBaseConsumption[-c(1,
                          2)])[colSums(dataBaseConsumption[dataBaseConsumption$Countr
                          input$Country,
                          -c(1,
                            2)]) !=
                          0], c(5:6)]),
                        sep = ": ")))))
            })
        })

    # Contamination data

    # Validate if mean level of
    # contamination is numeric
    meanCont <- reactive({
        validate(need(ifelse(!is.null(input$meanCont),
            !is.na(as.numeric(input$meanCont))),
            "Error: Mean level of contamination should be numeric"))
        as.numeric(input$meanCont)
    })
    # Validate if mean chance of
    # contamination is numeric
    meanChanceCont <- reactive({
        validate(need(ifelse(!is.null(input$meanChanceCont),
            !is.na(as.numeric(input$meanChanceCont))),
            "Error: Mean chance of contamination should be numeric"))
        as.numeric(input$meanChanceCont)
    })
    # Validate if SD level of
    # contamination is numeric
```

```
    sdCont <- reactive({
        validate(need(ifelse(!is.null(input$sdCont),
            !is.na(as.numeric(input$sdCont))),
            "Error: SD level of contamination should be numeric"))
        as.numeric(input$sdCont)
    })
    # Validate if SD chance of
    # contamination is numeric when
    # the box is checked
    sdChanceCont <- reactive({
        if (input$sdChanceContBox) {
            validate(need(ifelse(!is.null(input$sdChanceCont),
                !is.na(as.numeric(input$sdChanceCont))),
                "Error: SD chance of contamination should be numeric"))
            as.numeric(input$sdChanceCont)
        }
    })

    # Validate if nb points level
    # of contamination is numeric
    NbCont <- reactive({
        validate(need(ifelse(!is.null(input$NbCont),
            !is.na(as.numeric(input$NbCont))),
            "Error: Number of points for level of contamination should be numeric
        as.numeric(input$NbCont)
    })
    # Validate if nb points chance
    # of contamination is numeric
    # when the box is checked
    NbChanceCont <- reactive({
        if (input$sdChanceContBox) {
            validate(need(ifelse(!is.null(input$NbChanceCont),
                !is.na(as.numeric(input$NbChanceCont))),
                "Error: Number of points for chance of contamination should be nu
            as.numeric(input$NbChanceCont)
        }
    })
    # Hide or show the sd chance of
    # contamination input box
    observe({
        if (input$sdChanceContBox) {
            shinyjs::show("sdChanceCont")
            shinyjs::show("NbChanceCont")
```

```
        } else {
            shinyjs::hide("sdChanceCont")
            shinyjs::hide("NbChanceCont")
        }
    })

    # Calculate risk Change
    # calulate button status when
    # allergen and population are
    # selected
    observe({
        shinyjs::toggleState("update",
            ifelse(input$sdChanceContBox,
                input$NbCont !=
                  "" && input$meanCont !=
                  "" && input$meanChanceCont !=
                  "" && input$sdCont !=
                  "" && input$sdChanceCont !=
                  "" && input$NbChanceCont !=
                  "" && input$SelectAllergen !=
                  "" && input$SelectPopulation !=
                  "" && input$Country !=
                  "" && input$SelectGroup !=
                  "" && input$SelectItm !=
                  "" && is.null(datasetInput) ==
                  F, input$NbCont !=
                  "" && input$meanChanceCont !=
                  "" && input$sdCont !=
                  "" && input$SelectAllergen !=
                  "" && input$SelectPopulation !=
                  "" && input$Country !=
                  "" && input$SelectGroup !=
                  "" && input$SelectItm !=
                  "" && is.null(datasetInput) ==
                  F))
    })

    observeEvent(input$update,
        {

            risktotal = NULL
            riskGender = NULL
            # Select the data (population
```

```
            # and allergen) that will be
            # used for fitting the
            # challenge data
            if (input$SelectPopulation !=
                "All") {
                dataModel = dataChallenge()[dataChallenge()[,
                  names(dataChallenge()) ==
                    input$Allergen] ==
                  input$SelectAllergen &
                  dataChallenge()[,
                    names(dataChallenge()) ==
                      input$Population] ==
                    input$SelectPopulation,
                  ]
            } else {
                dataModel = dataChallenge()[dataChallenge()[,
                  names(dataChallenge()) ==
                    input$Allergen] ==
                  input$SelectAllergen,
                  ]
            }

            # Calculate summary statistics
            # consumption for the selected
            # food items and per gender
            if (input$SelectItm !=
                "Group") {
                dataset = subset(datasetInput(),
                  select = c("personid",
                    strsplit(input$SelectItm,
                      ": ")[[1]][1]))
            } else dataset = subset(datasetInput(),
                select = isolate({
                  c("personid",
                    as.character(selectGroup()[selectGroup()[,
                      5] %in% colnames(dataBaseConsumption[-c(1,
                      2)])][colSums(dataBaseConsumption[dataBaseConsumption$Countr
                      input$Country,
                      -c(1, 2)]) !=
                      0], 5]))
                }))
            joindata = dplyr::inner_join(PopulationInformation,
                dataset, by = "personid")
```

```
joindata[joindata ==
    0] <- NA


mdata <- melt(joindata,
    id = c("personid",
      "Country", "Sampling.weights",
      "Population",
      "Gender", "Age"))
mdata = mdata[, -which(colnames(mdata) ==
    "variable")]


totalsummary = mdata %>%
    summarise(N = sum(!is.na(value)),
      Mean = wtd.mean(value,
        Sampling.weights,
        na.rm = T),
      SD = sqrt(wtd.var(value,
        Sampling.weights,
        na.rm = T)))
gendersummary = mdata %>%
    group_by(Gender) %>%
    summarise(N = sum(!is.na(value)),
      Mean = wtd.mean(value,
        Sampling.weights,
        na.rm = T),
      SD = sqrt(wtd.var(value,
        Sampling.weights,
        na.rm = T))) %>%
    ungroup()
gendersummary$Gender = as.character(gendersummary$Gender)
summaryCons = rbind(c(Gender = "All",
    totalsummary),
    gendersummary)
# Calculates mean and sd chance
# of consumption
if (dim(joindata)[2] >
    7) {
    chanceCons = colSums(replicate(n = 1000,
      sample(rowSums(!is.na(joindata[,
        -c(1:6)])) !=
        0, dim(joindata)[1],
```

```
                replace = T,
                prob = joindata$Sampling.weights),
              simplify = T))/dim(joindata)[1]
        } else {
            chanceCons = colSums(replicate(n = 1000,
              sample(!is.na(joindata[,
                -c(1:6)]) !=
                0, dim(joindata)[1],
                replace = T,
                prob = joindata$Sampling.weights),
              simplify = T))/dim(joindata)[1]

        }
        consumptionChance = function(joindata) {
            if (dim(joindata)[2] >
              7) {
              return(colSums(replicate(n = 1000,
                sample(rowSums(!is.na(joindata[,
                  -c(1:6)])) !=
                  0, dim(joindata)[1],
                  replace = T,
                  prob = joindata$Sampling.weights),
                simplify = T))/dim(joindata)[1])
            } else {
              return(chanceCons = colSums(replicate(n = 1000,
                sample(!is.na(joindata[,
                  -c(1:6)]) !=
                  0, dim(joindata)[1],
                  replace = T,
                  prob = joindata$Sampling.weights),
                simplify = T))/dim(joindata)[1])
            }
        }
        chanceConsGender = do.call(rbind,
            lapply(split(joindata,
              joindata$Gender),
              consumptionChance))
        summaryChance = rbind(c(Gender = "All",
            mean(chanceCons),
            sd(chanceCons)),
            cbind(Gender = rownames(chanceConsGender),
              apply(chanceConsGender,
                1, mean), apply(chanceConsGender,
```

```
            1, sd)))
        summaryChance = as.data.frame(summaryChance)
        summaryChance[, 2] = as.numeric(as.character(summaryChance[,
            2]))
        summaryChance[, 2] = round(summaryChance[,
            2] * 100, 2)
        colnames(summaryChance)[2] = "Mean Consumption Chance"
        summaryChance[, 3] = as.numeric(as.character(summaryChance[,
            3]))
        summaryChance[, 3] = round(summaryChance[,
            3] * 100, 2)
        colnames(summaryChance)[3] = "SD Consumption Chance"
        output$summaryCons <- renderTable({
            isolate({
              summaryCons
            })
        }, include.rownames = FALSE)
        output$summaryChance <- renderTable({
            isolate({
              summaryChance
            })
        }, include.rownames = FALSE)

        # Risk

        # Risk calculated for Male and
        # female
        risktotal = risk(consumption = isolate({
            joindata
        }), MeanContamination = isolate({
            meanCont()
        }), SDContamination = isolate({
            sdCont()
        }), NbPointContamination = isolate({
            NbCont()
        }), meanChanceCont = isolate({
            meanChanceCont()
        }), SDChanceCont = ifelse(input$sdChanceContBox,
            sdChanceCont(),
            NA), NbPointChanceCont = ifelse(input$sdChanceContBox,
            NbChanceCont(),
            NA), challenge = isolate({
            dataModel
```

```
            }), NOAEL = isolate({
                input$NOAEL
            }), LOAEL = isolate({
                input$LOAEL
            }), meanPrevalence = isolate({
                prevalence[prevalence$Allergen ==
                  input$SelectAllergen &
                  prevalence$Country ==
                    input$Country &
                  prevalence$Population ==
                    input$SelectPopulation,
                  4]
            }), SDPrevalence = isolate({
                prevalence[prevalence$Allergen ==
                  input$SelectAllergen &
                  prevalence$Country ==
                    input$Country &
                  prevalence$Population ==
                    input$SelectPopulation,
                  5]
            }), NbPrevalence = isolate({
                prevalence[prevalence$Allergen ==
                  input$SelectAllergen &
                  prevalence$Country ==
                    input$Country &
                  prevalence$Population ==
                    input$SelectPopulation,
                  6]
            }))
            # Risk calculated by gender
            riskGender = lapply(split(isolate({
                joindata
            }), joindata$Gender),
                risk, MeanContamination = isolate({
                  meanCont()
                }), SDContamination = isolate({
                  sdCont()
                }), NbPointContamination = isolate({
                  NbCont()
                }), meanChanceCont = isolate({
                  meanChanceCont()
                }), SDChanceCont = ifelse(input$sdChanceContBox,
                  sdChanceCont(),
```

```
            NA), NbPointChanceCont = ifelse(input$sdChanceContBox,
            NbChanceCont(),
            NA), challenge = isolate({
            dataModel
        }), NOAEL = isolate({
            input$NOAEL
        }), LOAEL = isolate({
            input$LOAEL
        }), meanPrevalence = isolate({
            prevalence[prevalence$Allergen ==
              input$SelectAllergen &
              prevalence$Country ==
                input$Country &
              prevalence$Population ==
                input$SelectPopulation,
              4]
        }), SDPrevalence = isolate({
            prevalence[prevalence$Allergen ==
              input$SelectAllergen &
              prevalence$Country ==
                input$Country &
              prevalence$Population ==
                input$SelectPopulation,
              5]
        }), NbPrevalence = isolate({
            prevalence[prevalence$Allergen ==
              input$SelectAllergen &
              prevalence$Country ==
                input$Country &
              prevalence$Population ==
                input$SelectPopulation,
              6]
        }))
    # Output risk table and title
    show("results")
    output$title1 <- renderText({
        isolate({
          paste0("Allergen=",
            input$SelectAllergen,
            ", Population=",
            input$SelectPopulation)
        })
    })
```

```
output$title2 <- renderText({
    isolate({
      paste0("Country=",
        input$Country,
        ", Group=",
        input$SelectGroup,
        ", Food Item=",
        input$SelectItm)
    })
})
output$title3 <- renderText({
    isolate({
      paste0("Mean contamination level=",
        meanCont(),
        ", SD contamination level=",
        sdCont(), ", Number of points contamination level=",
        NbCont())
    })
})
output$title4 <- renderText({
    isolate({
      ifelse(input$sdChanceContBox,
        paste0("Mean contamination chance=",
          meanChanceCont(),
          ", SD contamination chance=",
          sdChanceCont(),
          ", Number of points contamination chance=",
          NbChanceCont()),
        paste0("Mean contamination chance=",
          meanChanceCont()))
    })
})

output$RiskTable <- renderTable(risktotal$table)
output$RiskTableMale <- renderTable(riskGender$Male$table)
output$RiskTableFemale <- renderTable(riskGender$Female$table)

# Show dowloadData button when
# the risk are calculated
observeEvent(!is.null(risktotal) &&
    !is.null(riskGender),
    {
      show("downloadData")
```

```
    })
  # Shaping the file for
  # exportation when dowloadDat
  # is clicked
  overall = cbind.data.frame(Overall = as.character(rownames(risktotal$save
      risktotal$saveFile)
  overall$Overall = as.character(overall$Overall)
  male = cbind.data.frame(Overall = as.character(rownames(riskGender$Male$s
      riskGender$Male$saveFile)
  male$Overall = as.character(male$Overall)
  female = cbind.data.frame(Overall = as.character(rownames(riskGender$Fema
      riskGender$Female$saveFile)
  female$Overall = as.character(female$Overall)

  exportFile = do.call("rbind.data.frame",
      list(overall, c("Male",
        colnames(riskGender$Male$saveFile)),
        male, c("Female",
          colnames(riskGender$Female$saveFile)),
        female))
  output$downloadData <- downloadHandler(filename = "riskTable.csv",
      content = function(file) {
        write.csv(exportFile,
          file, row.names = FALSE)
      })

  # Plotting threshold VS
  # consumption when allergic
  # reactions occur
  dataPlot = cbind.data.frame(threshold = c(risktotal$simulation$threshold)
      consumption = c(risktotal$simulation$consumption),
      reaction = c(risktotal$simulation$reaction))
  dataPlot = subset(dataPlot,
      reaction == T)

  consPlot = mdata %>%
      summarise(Mean = wtd.mean(value,
        Sampling.weights,
        na.rm = T), P90 = wtd.quantile(value,
        Sampling.weights,
        probs = 0.9,
        na.rm = TRUE))
  names(consPlot) = c("Average consumption",
```

```
        "p90 consumption")
    y <- cbind.data.frame(Summary = names(consPlot),
        Intercept = t(consPlot))
    z <- cbind.data.frame(Summary = c("ED05 (Log-Normal)",
        "ED10 (Log-Normal)",
        "Lowest LOAEL"),
        Intercept = c(risktotal$ED[2:3],
          min(dataModel[,
            names(dataModel) ==
              input$LOAEL],
            na.rm = T)))
    output$reactionPlot <- renderPlot(ggplot(dataPlot,
        aes(x = threshold,
          y = consumption)) +
        geom_point() +
        xlab("Individual Threshold Simulated (mg)") +
        ylab("Consumption Amount (g)") +
        theme_bw() + theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank()) +
        geom_hline(aes(yintercept = Intercept,
          colour = Summary),
          data = y, show_guide = TRUE,
          size = 1) + geom_vline(aes(xintercept = Intercept,
        colour = Summary),
        data = z, show_guide = TRUE,
        size = 1))

    # MS Risk

    # Change calulate button status
    # when allergen and population
    # are selected
    observe({
        shinyjs::toggleState("updateMS",
          input$MSproducts !=
            "" && input$percProducts !=
            "")
    })
    # Validate if product market
    # share is numeric
    MSproducts <- reactive({
        validate(need(ifelse(!is.null(input$MSproducts),
          !is.na(as.numeric(input$MSproducts))),
```

```
                  "Error: Product market share should be numeric"))
              as.numeric(input$MSproducts)/100
          })
          # Validate if number of
          # products is numeric
          percProducts <- reactive({
              validate(need(ifelse(!is.null(input$percProducts),
                !is.na(as.numeric(input$percProducts))),
                "Error: Percentage of products/packages/lots implicated should be n
              as.numeric(input$percProducts)
          })
          # Output table with number of
          # allaergic reaction per
          # population
          observeEvent(input$updateMS,
              {
                colnames(risktotal$saveFile[,
                  c(1, 4, 7)]) = c("Log-Normal",
                  "Log-Logistic",
                  "Weibull")
                output$RiskMS <- renderTable({
                  MSproducts() *
                    percProducts() *
                    risktotal$saveFile[,
                      c(1, 4,
                        7)]
                }, digits = 0)
              })
      })
})
```

C.4 R code: risk function

```
risk = function(consumption, MeanContamination,
    SDContamination, NbPointContamination,
    meanChanceCont, SDChanceCont = NA,
    NbPointChanceCont = NA, challenge,
    NOAEL, LOAEL, meanPrevalence,
    SDPrevalence, NbPrevalence,
    N = 10000, K = 1000) {
```

```
    # consumption input: data.frame
    # with 7 or more columns (first
    # 6 columns= information about
    # consumers, following columns=
    # food items consumption)
    # MeanContamination: numeric,
    # mean contamination
    # SDContamination: numeric, SD
    # contamination
    # NbPointContamination:
    # numeric, number of points on
    # which the mean and SD
    # contamination was measured
    # meanChanceCont: numeric, mean
    # chance of contamination
    # SDChanceCont: numeric,
    # optional, SD chance of
    # contamination
    # NbPointChanceCont: numeric,
    # optional, number of points on
    # which the mean and SD chance
    # of contamination was measured
    # challenge: data.frame, the
    # challenge data with threshold
    # NOAEL: character, name of the
    # columns that contains the
    # NOAEL values LOAEL:
    # character, name of the
    # columns that contains the
    # LOAEL values meanPrevalence:
    # numeric, mean prevalence of
    # allergic reaction in the
    # population SDPrevalence:
    # numeric, SD prevalence of
    # allergic reaction in the
    # population NbPrevalence:
    # numeric, number of points on
    # which the mean and SD
    # prevalence of allergic
    # reaction in the population
    # was measured

    library(survival)
```

```
library(parallel)
library(reshape)
library(eha)
cl = makeCluster(detectCores() -
    1)
clusterEvalQ(cl, library(eha))

# print(summary(consumption))
# print(MeanContamination)
# print(SDContamination)
# print(NbPointContamination)
# print(meanChanceCont)
# print(SDChanceCont)
# print(NbPointChanceCont)
# print(summary(challenge))
# print(NOAEL) print(LOAEL)
# print(meanPrevalence)
# print(SDPrevalence)
# print(NbPrevalence) joindata
# Melt the data, so there is
# only 1 column with
# consumption values
mdata <- melt(consumption,
    id = c("personid", "Country",
        "Sampling.weights",
        "Population", "Gender",
        "Age"))
# print(summary(mdata)) Remove
# column with the name of the
# food items that is not used
# later
mdata = mdata[, -which(colnames(mdata) ==
    "variable")]
# Only keeps consumption values
# that are not missing
mdata = mdata[which(!is.na(mdata$value)),
    ]
# Export objects for parallel
# computing
# clusterExport(cl,c('mdata','N','K','consumption'),envir
# = environment()) Sampling in
# empirical distribution with
# remplacment according to
```

```
    # survey's weights
    # print(class(mdata))
    cons = parSapply(cl, 1:K, function(i) {
        sample(mdata[, 7], N, replace = T,
            prob = mdata[, 3])
    }, simplify = T)
    # cons=replicate(K,sample(mdata[,7],N,replace=T,prob=mdata[,3]),simplify
    # = T)

    # dim(consumption)#[1] N=10000
    # K=1000

    # Calculates chance of
    # consumption with bootstrap
    # (with surveys weights)
    if (dim(consumption)[2] > 7) {
        chanceCons = colSums(parSapply(cl,
            1:K, function(i) {
                sample(rowSums(!is.na(consumption[,
                  !colnames(consumption) %in%
                    c("personid",
                      "Country",
                      "Sampling.weights",
                      "Population",
                      "Gender",
                      "Age")])) !=
                  0, dim(consumption)[1],
                  replace = T,
                  prob = consumption$Sampling.weights)
            }, simplify = T))/dim(consumption)[1]

        # chanceCons=colSums(replicate(K,sample(rowSums(!is.na(consumption[,!coln
        # ]))!=0,dim(consumption)[1],replace=T,
        # prob=consumption$Sampling.weights),simplify
        # = T))/dim(consumption)[1]
    } else {
        chanceCons = colSums(parSapply(cl,
            1:K, function(i) {
                sample(!is.na(consumption[,
                  !colnames(consumption) %in%
                    c("personid",
                      "Country",
                      "Sampling.weights",
```

```
                    "Population",
                    "Gender",
                    "Age")]) !=
                0, dim(consumption)[1],
                replace = T,
                prob = consumption$Sampling.weights)
        }, simplify = T))/dim(consumption)[1]

    # chanceCons=colSums(replicate(K,sample(!is.na(consumption[,!colnames(consump
    # ])!=0,dim(consumption)[1],replace=T,
    # prob=consumption$Sampling.weights),simplify
    # = T))/dim(consumption)[1]
}

# Contamination simulation with
# probability from the vector
# of chance of contamination
binomialRandom = function(proba,
    N) {
    rbinom(n = N, size = 1,
        prob = proba)
}
consumptionChance = mcmapply(binomialRandom,
    chanceCons, N)
# consumptionChance=mapply(binomialRandom,chanceCons,N)

# dim(consumptionChance)#[1]
# N=10000 K=1000

# Challenge data Formula
# regression
formulaRegression = as.formula(paste("Surv(time=",
    NOAEL, ", time2=", LOAEL,
    ", type='interval2')~1"))
# print(formulaRegression) Fit
# the 3 distributions to
# challenge data and extract
# the standard deviation of the
# parameters
LogNormalReg <- survreg(formulaRegression,
    data = challenge, na.action = na.omit,
    dist = "lognormal")
SEsLogNorm = sqrt(diag(summary(LogNormalReg)$var))
```

```
LogLogisticReg <- survreg(formulaRegression,
    data = challenge, na.action = na.omit,
    dist = "loglogistic")
SEsLogLogistic = sqrt(diag(summary(LogLogisticReg)$var))
WeibullReg <- survreg(formulaRegression,
    data = challenge, na.action = na.omit,
    dist = "weibull")
SEsWeibull = sqrt(diag(summary(WeibullReg)$var))

# Threshold LogNormal
logNormRandom = function(mean,
    sd, N) {
    rlnorm(N, meanlog = mean,
        sdlog = sd)
}
LogNormalResponse = mcmapply(logNormRandom,
    rnorm(K, LogNormalReg$coeff,
        SEsLogNorm[1]), exp(rnorm(K,
        log(LogNormalReg$scale),
        SEsLogNorm[2])), N)
# LogNormalResponse=mapply(logNormRandom,rnorm(K,LogNormalReg$coeff,SEsLogNor
# dim(LogNormalResponse) #[1]
# 10000 1000 Threshold Weibull
WeibullRandom = function(scale,
    shape, N) {
    rweibull(N, shape = shape,
        scale = scale)
}
WeibullResponse = mcmapply(WeibullRandom,
    exp(rnorm(K, WeibullReg$coeff,
        SEsWeibull[1])), 1/exp(rnorm(K,
        log(WeibullReg$scale),
        SEsWeibull[2])), N)
# WeibullResponse=mapply(WeibullRandom,exp(rnorm(K,WeibullReg$coeff,SEsWeibul

# dim(WeibullResponse) #[1]
# 10000 1000 Threshold
# LogLogistic
LogLogisticRandom = function(scale,
    shape, N) {
    rllogis(N, shape = shape,
        scale = scale)
}
```

```
    LogLogisticResponse = mcmapply(LogLogisticRandom,
        exp(rnorm(K, LogLogisticReg$coeff,
            SEsLogLogistic[1])),
        1/exp(rnorm(K, log(LogLogisticReg$scale),
            SEsLogLogistic[2])),
        N)
    # LogLogisticResponse=mapply(LogLogisticRandom,exp(rnorm(K,LogLogisticReg$coeff,S
    # dim(LogLogisticResponse) #[1]
    # 10000 1000

    # Prevalence Simulating K
    # prevalence value from the
    # publications information,
    # mean, sd and number of points
    prevalenceVec = rnorm(K, mean = meanPrevalence,
        sd = SDPrevalence/sqrt(NbPrevalence))
    # prevalenceVec[prevalenceVec<0]=0
    while (sum(prevalenceVec <
        0) != 0) {
        prevalenceVec[prevalenceVec <
            0] = rnorm(sum(prevalenceVec <
            0), mean = meanPrevalence,
            sd = SDPrevalence/sqrt(NbPrevalence))
    }
    # Prevalence from binomial
    # distribution with probaility
    # from the vector of prevalence
    # previously simulated
    prevalence = mcmapply(binomialRandom,
        prevalenceVec, N)
    # prevalence=mapply(binomialRandom,prevalenceVec,N)

    # dim(prevalence)#[1] N=10000
    # K=1000

    # Level of Contamination
    mm2 = log((MeanContamination^2)/sqrt(MeanContamination^2 +
        SDContamination^2))
    ss2 = sqrt(log(1 + (SDContamination/MeanContamination)^2))
    contMean = rnorm(K, mean = mm2,
        sd = ss2/sqrt(NbPointContamination))
    contSD = sqrt(ss2^2 * rchisq(K,
        NbPointContamination -
```

```
        1)/(NbPointContamination -
    1))
concentration = mcmapply(logNormRandom,
    contMean, contSD, N)
# concentration=mapply(logNormRandom,contMean,contSD,N)

# dim(concentration)#[1] 10000
# 1000

# Chance of contamination
if (!is.na(SDChanceCont) &&
    !is.na(NbPointChanceCont)) {
    chanceCont = rnorm(K, mean = meanChanceCont,
        sd = SDChanceCont/sqrt(NbPointChanceCont))
    # chanceCont[chanceCont<0]=0
    while (sum(chanceCont <
        0) != 0) {
        chanceCont[chanceCont <
            0] = rnorm(sum(chanceCont <
            0), mean = meanChanceCont,
            sd = SDChanceCont/sqrt(NbPointChanceCont))
    }

}
contaminationChance = mcmapply(binomialRandom,
    if (is.na(SDChanceCont) &&
        is.na(NbPointChanceCont))
        rep(meanChanceCont,
            K) else chanceCont, N)
# contaminationChance=mapply(binomialRandom,if(is.na(SDChanceCont)&&is.na(NbP
# rep(meanChanceCont,K) else
# chanceCont,N)

# dim(contaminationChance)#[1]
# N=10000 K=1000
stopCluster(cl)
# User risk
exposureUser = contaminationChance *
    concentration * cons *
    10^-3
reactionUserLogNormal = exposureUser >
    LogNormalResponse
riskUserLogNormal = colSums(reactionUserLogNormal)/N
```

```
    reactionUserWeibull = exposureUser >
        WeibullResponse
    riskUserWeibull = colSums(reactionUserWeibull)/N
    reactionUserLogLogistic = exposureUser >
        LogLogisticResponse
    riskUserLogLogistic = colSums(reactionUserLogLogistic)/N

    # Allergic population risk
    exposureAllergic = exposureUser *
        consumptionChance
    reactionAllergicLogNormal = exposureAllergic >
        LogNormalResponse
    riskAllergicLogNormal = colSums(reactionAllergicLogNormal)/N
    reactionAllergicWeibull = exposureAllergic >
        WeibullResponse
    riskAllergicWeibull = colSums(reactionAllergicWeibull)/N
    reactionAllergicLogLogistic = exposureAllergic >
        LogLogisticResponse
    riskAllergicLogLogistic = colSums(reactionAllergicLogLogistic)/N

    # Overall population risk
    exposureOverall = exposureAllergic *
        prevalence
    reactionOverallLogNormal = exposureOverall >
        LogNormalResponse
    riskOverallLogNormal = colSums(reactionOverallLogNormal)/N
    reactionOverallWeibull = exposureOverall >
        WeibullResponse
    riskOverallWeibull = colSums(reactionOverallWeibull)/N
    reactionOverallLogLogistic = exposureOverall >
        LogLogisticResponse
    riskOverallLogLogistic = colSums(reactionOverallLogLogistic)/N

    # Summary table with risks
    summaryTable = data.frame(matrix(nrow = 3,
        ncol = 3))
    rownames(summaryTable) = c("User Risk",
        "Allergic Population",
        "Overall Population")
    colnames(summaryTable) = c("Log-Normal",
        "Log-Logistic", "Weibull")
    summaryTable[1, 1] = paste0(round(mean(riskUserLogNormal) *
        100, 2), "% (", round(t.test(riskUserLogNormal)$conf.int[1] *
```

```
        100, 2), "-", round(t.test(riskUserLogNormal)$conf.int[2] *
        100, 2), ")")
    summaryTable[1, 2] = paste0(round(mean(riskUserLogLogistic) *
        100, 2), "% (", round(t.test(riskUserLogLogistic)$conf.int[1] *
        100, 2), "-", round(t.test(riskUserLogLogistic)$conf.int[2] *
        100, 2), ")")
    summaryTable[1, 3] = paste0(round(mean(riskUserWeibull) *
        100, 2), "% (", round(t.test(riskUserWeibull)$conf.int[1] *
        100, 2), "-", round(t.test(riskUserWeibull)$conf.int[2] *
        100, 2), ")")
    summaryTable[2, 1] = paste0(round(mean(riskAllergicLogNormal) *
        100, 3), "% (", round(t.test(riskAllergicLogNormal)$conf.int[1] *
        100, 3), "-", round(t.test(riskAllergicLogNormal)$conf.int[2] *
        100, 3), ")")
    summaryTable[2, 2] = paste0(round(mean(riskAllergicLogLogistic) *
        100, 3), "% (", round(t.test(riskAllergicLogLogistic)$conf.int[1] *
        100, 3), "-", round(t.test(riskAllergicLogLogistic)$conf.int[2] *
        100, 3), ")")
    summaryTable[2, 3] = paste0(round(mean(riskAllergicWeibull) *
        100, 3), "% (", round(t.test(riskAllergicWeibull)$conf.int[1] *
        100, 3), "-", round(t.test(riskAllergicWeibull)$conf.int[2] *
        100, 3), ")")
    summaryTable[3, 1] = paste0(round(mean(riskOverallLogNormal) *
        100, 4), "% (", round(t.test(riskOverallLogNormal)$conf.int[1] *
        100, 4), "-", round(t.test(riskOverallLogNormal)$conf.int[2] *
        100, 4), ")")
    summaryTable[3, 2] = paste0(round(mean(riskOverallLogLogistic) *
        100, 4), "% (", round(t.test(riskOverallLogLogistic)$conf.int[1] *
        100, 4), "-", round(t.test(riskOverallLogLogistic)$conf.int[2] *
        100, 4), ")")
    summaryTable[3, 3] = paste0(round(mean(riskOverallWeibull) *
        100, 4), "% (", round(t.test(riskOverallWeibull)$conf.int[1] *
        100, 4), "-", round(t.test(riskOverallWeibull)$conf.int[2] *
        100, 4), ")")

    # Summary table with risks
    saveFile = data.frame(matrix(nrow = 3,
        ncol = 9))
    rownames(saveFile) = c("User Risk",
        "Allergic Population",
        "Overall Population")
    colnames(saveFile) = c("Log-Normal: Mean",
        "Log-Normal: CI-", "Log-Normal: CI+",
```

```
        "Log-Logistic: Mean", "Log-Logistic: CI-",
        "Log-Logistic: CI+", "Weibull: Mean",
        "Weibull: CI-", "Weibull: CI+")
    saveFile[1, ] = c(mean(riskUserLogNormal),
        t.test(riskUserLogNormal)$conf.int[1],
        t.test(riskUserLogNormal)$conf.int[2],
        mean(riskUserLogLogistic),
        t.test(riskUserLogLogistic)$conf.int[1],
        t.test(riskUserLogLogistic)$conf.int[2],
        mean(riskUserWeibull),
        t.test(riskUserWeibull)$conf.int[1],
        t.test(riskUserWeibull)$conf.int[2])
    saveFile[2, ] = c(mean(riskAllergicLogNormal),
        t.test(riskAllergicLogNormal)$conf.int[1],
        t.test(riskAllergicLogNormal)$conf.int[2],
        mean(riskAllergicLogLogistic),
        t.test(riskAllergicLogLogistic)$conf.int[1],
        t.test(riskAllergicLogLogistic)$conf.int[2],
        mean(riskAllergicWeibull),
        t.test(riskAllergicWeibull)$conf.int[1],
        t.test(riskAllergicWeibull)$conf.int[2])
    saveFile[3, ] = c(mean(riskOverallLogNormal),
        t.test(riskOverallLogNormal)$conf.int[1],
        t.test(riskOverallLogNormal)$conf.int[2],
        mean(riskOverallLogLogistic),
        t.test(riskOverallLogLogistic)$conf.int[1],
        t.test(riskOverallLogLogistic)$conf.int[2],
        mean(riskOverallWeibull),
        t.test(riskOverallWeibull)$conf.int[1],
        t.test(riskOverallWeibull)$conf.int[2])
    saveFile = saveFile * 100

    # summaryTable
    output = list()
    output$table = summaryTable
    output$saveFile = saveFile
    output$simulation = list(consumption = cons,
        threshold = LogNormalResponse,
        reaction = reactionUserLogNormal)
    output$ED = unique(predict(LogNormalReg,
        type = "quantile", p = c(0.01,
            0.05, 0.1, 0.5), se = T)[[1]])
    return(output)
```

```
}
```

Bibliography

[1] Annual Report of Incidents, 2014.

[2] T W Anderson. On the Distribution of the Two-Sample Cramer-von Mises Criterion. pages 1148–1159, 1962.

[3] J Barnett, J Leftwich, K Muncer, K Grimshaw, R Shepherd, M M Raats, M H Gowland, and J S Lucas. How do peanut and nut-allergic consumers use information on the packaging to avoid allergens? *Allergy: European Journal of Allergy and Clinical Immunology*, 66(7):969–978, 2011.

[4] Julie Barnett, Kate Muncer, Jo Leftwich, Richard Shepherd, Monique M Raats, M Hazel Gowland, Kate Grimshaw, and Jane S Lucas. Using 'may contain' labelling to inform food choice: a qualitative study of nut allergic consumers. *BMC public health*, 11(1):734, 2011.

[5] W. J. Conover. Practical Nonparametric Statistics., 1971.

[6] R W R Crevel, B K Ballmer-Weber, T Holzhauser, J O B Hourihane, A C Knulst, A R Mackie, F Timmermans, and S L Taylor. Thresholds for food allergens and their value to different stakeholders. *Allergy: European Journal of Allergy and Clinical Immunology*, 63(5):597–609, 2008.

[7] Steven M Gendel. Comparison of international food allergen labeling regulations. *Regulatory Toxicology and Pharmacology*, 63(2):279–285, 2012.

[8] Richard Johnson. *Miller & Freund's probability and statistics for engineers.* Prentice Hall, Boston, 2011.

[9] C B Madsen, S Hattersley, K J Allen, K Beyer, C H Chan, S B Godefroy, R Hodgson, E N C Mills, a. Muñoz-Furlong, S Schnadt, R Ward, M Wickman, and R Crevel. Can we define a tolerable level of risk in food allergy?

Report from a EuroPrevall/UK Food Standards Agency workshop. *Clinical and Experimental Allergy*, 42(1):30–37, 2012.

[10] Charlotte Madsen, René Crevel, Clare Mills, and Steve Taylor. *Risk management for food allergy.* Academic Press, 2013.

[11] David Makowski. Uncertainty and sensitivity analysis in quantitative pest risk assessments; practical rules for risk assessors. *NeoBiota*, 18:157–171, 2013.

[12] B. I. Nwaru, L. Hickstein, S. S. Panesar, G. Roberts, A. Muraro, and A. Sheikh. Prevalence of common food allergies in Europe: A systematic review and meta-analysis. *Allergy: European Journal of Allergy and Clinical Immunology*, 69(8):992–1007, 2014.

[13] Loup Rimbaud, Fanny Heraud, Sébastien La Vieille, Jean Charles Leblanc, and Amélie Crepet. Quantitative risk assessment relating to adventitious presence of allergens in food: A probabilistic model applied to peanut in chocolate. *Risk Analysis*, 30(1):7–19, 2010.

[14] M. Q I Spanjersberg, a. G. Kruizinga, M. a J Rennen, and G. F. Houben. Risk assessment and food allergy: the probabilistic model applied to allergens. *Food and Chemical Toxicology*, 45(1):49–54, 2007.

[15] The Anaphylaxis Campaign. 'May Contain' labelling - the consumer's perspective. pages —-, 2002.

[16] Paul J Turner, Andrew Kemp, and Dianne Campbell. Advisory food labels: consumers with allergies need more than "traces" of information. *British Medical Journal*, 343(d6180):—-, 2011.

[17] R Ward, R Crevel, I Bell, N Khandke, C Ramsay, and S Paine. A vision for allergen management best practice in the food industry. *Trends in Food Science and Technology*, 21(12):619–625, 2010.